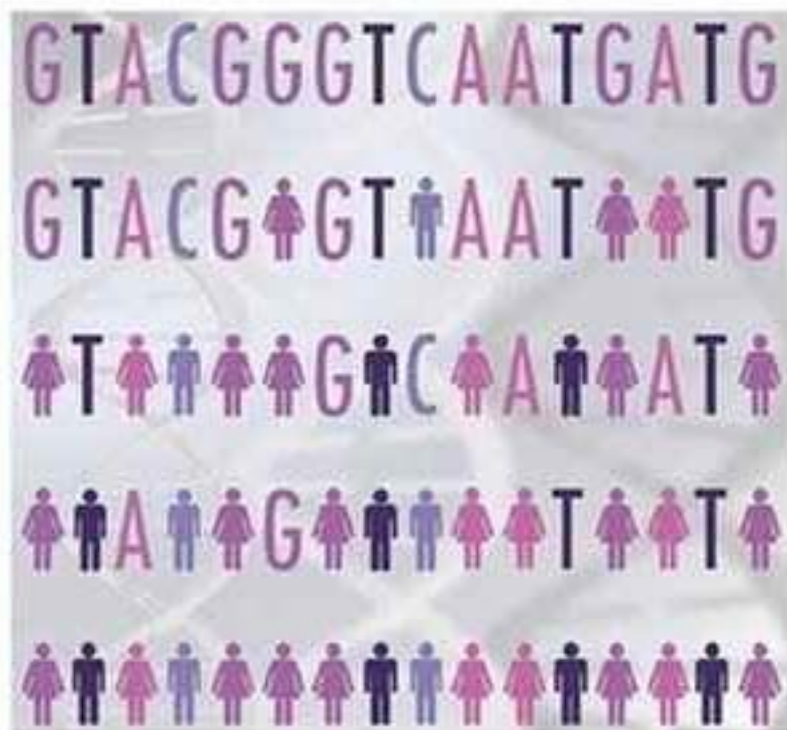


Edited by Michal Janitz

WILEY-VCH

Next-Generation Genome Sequencing

Towards Personalized Medicine



**Next-Generation
Genome Sequencing**

*Edited by
Michal Janitz*

Related Titles

Dehmer, M., Emmert-Streib, F. (eds.)

Analysis of Microarray Data

2008

ISBN: 978-3-527-31822-3

Helms, V.

Principles of Computational Cell Biology

2008

ISBN: 978-3-527-31555-1

Knudsen, S.

Cancer Diagnostics with DNA Microarrays

2006

ISBN: 978-0-471-78407-4

Sensen, C. W. (ed.)

Handbook of Genome Research

Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues

2005

ISBN: 978-3-527-31348-8

Next-Generation Genome Sequencing

Towards Personalized Medicine

Edited by
Michal Janitz



**WILEY-
BLACKWELL**

WILEY-VCH Verlag GmbH & Co. KGaA

The Editor

Dr. Michal Janitz
Max Planck Institute for
Molecular Genetics
Fabeckstr. 60-62
14195 Berlin
Germany

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek

Die Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA,
Weinheim

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Composition Thomson Digital, Noida, India

Printing Betz-Druck GmbH, Darmstadt

Bookbinding Litges & Dopf GmbH, Heppenheim

Printed in the Federal Republic of Germany

Printed on acid-free paper

ISBN: 978-3-527-32090-5

Contents

Preface XIII

List of Contributors XVII

Part One Sanger DNA Sequencing 1

1 Sanger DNA Sequencing 3

Artem E. Men, Peter Wilson, Kirby Siemering, and Susan Forrest

- 1.1 The Basics of Sanger Sequencing 3
- 1.2 Into the Human Genome Project (HGP) and Beyond 6
- 1.3 Limitations and Future Opportunities 7
- 1.4 Bioinformatics Holds the Key 8
- 1.5 Where to Next? 9
- References 10

Part Two Next-Generation Sequencing: Toward Personalized Medicine 13

2 Illumina Genome Analyzer II System 15

Abizar Lakdawalla and Harper VanSteenhouse

- 2.1 Library Preparation 15
- 2.2 Cluster Creation 17
- 2.3 Sequencing 19
- 2.4 Paired End Reads 19
- 2.5 Data Analysis 20
- 2.6 Applications 21
 - 2.6.1 Genome Sequencing Applications 23
 - 2.6.2 Epigenomics 23
 - 2.6.3 Transcriptome Analysis 23
 - 2.6.4 Protein–Nucleic Acid Interactions 26
 - 2.6.5 Multiplexing 26

2.7	Conclusions	26
	References	27

3 **Applied Biosystems SOLiD™ System: Ligation-Based Sequencing** 29

Vicki Pandey, Robert C. Nutter, and Ellen Prediger

3.1	Introduction	29
3.2	Overview of the SOLiD™ System	29
3.2.1	The SOLiD Platform	30
3.2.1.1	Library Generation	30
3.2.1.2	Emulsion PCR	31
3.2.1.3	Bead Purification	31
3.2.1.4	Bead Deposition	33
3.2.1.5	Sequencing by Ligation	33
3.2.1.6	Color Space and Base Calling	35
3.3	SOLiD™ System Applications	35
3.3.1	Large-Scale Resequencing	35
3.3.2	<i>De novo</i> Sequencing	35
3.3.3	Tag-Based Gene Expression	36
3.3.4	Whole Transcriptome Analysis	37
3.3.5	Whole Genome Resequencing	38
3.3.6	Whole Genome Methylation Analysis	38
3.3.7	Chromatin Immunoprecipitation	39
3.3.8	MicroRNA Discovery	39
3.3.9	Other Tag-Based Applications	40
3.4	Conclusions	40
	References	41

4 **The Next-Generation Genome Sequencing: 454/Roche GS FLX** 43

Lei Du and Michael Egholm

4.1	Introduction	43
4.2	Technology Overview	44
4.3	Software and Bioinformatics	47
4.3.1	Whole Genome Assembly	47
4.3.2	Resequencing and Mutation Detection	47
4.3.3	Ultradeep Sequencing	47
4.4	Research Applications	49
	References	51

5 **Polony Sequencing: History, Technology, and Applications** 57

Jeremy S. Edwards

5.1	Introduction	57
5.2	History of Polony Sequencing	57
5.2.1	Introduction to Polonies	58
5.2.2	Evolution of Polonies	59
5.2.3	Current Applications of the Original Polonies Method	61

5.3	Polony Sequencing	62
5.3.1	Constructing a Sequencing Library	63
5.3.2	Loading the Library onto Beads Using BEAMing	64
5.3.3	Immobilizing the Beads in the Sequencing Flow Cell	65
5.3.4	Sequencing	66
5.3.5	Data Analysis	68
5.4	Applications	69
5.4.1	Human Genome Sequencing	69
5.4.1.1	Requirements of an Ultrahigh-Throughput Sequencing Technology	69
5.4.2	Challenges of Sequencing the Human Genome with Short Reads	70
5.4.2.1	Chromosome Sequencing	72
5.4.2.2	Exon Sequencing	72
5.4.2.3	Impact on Medicine	72
5.4.3	Transcript Profiling	73
5.4.3.1	Polony SAGE	73
5.4.3.2	Transcript Characterization with Polony SAGE	73
5.4.3.3	Digital Karyotyping	75
5.5	Conclusions	75
	References	76

Part Three The Bottleneck: Sequence Data Analysis 77

6	Next-Generation Sequence Data Analysis	79
	<i>Leonard N. Bloksberg</i>	
6.1	Why Next-Generation Sequence Analysis is Different?	79
6.2	Strategies for Sequence Searching	80
6.3	What is a “Hit,” and Why it Matters for NGS?	82
6.3.1	Word Hit	82
6.3.2	Segment Hit	82
6.3.3	SeqID Hit or Gene Hit	82
6.3.4	Region Hit	82
6.3.5	Mapped Hit	83
6.3.6	Syntenic Hit	83
6.4	Scoring: Why it is Different for NGS?	83
6.5	Strategies for NGS Sequence Analysis	84
6.6	Subsequent Data Analysis	86
	References	87
7	DNASTAR’s Next-Generation Software	89
	<i>Tim Durfee and Thomas E. Schwei</i>	
7.1	Personalized Genomics and Personalized Medicine	89
7.2	Next-Generation DNA Sequencing as the Means to Personalized Genomics	89

7.3	Strengths of Various Platforms	90
7.4	The Computational Challenge	90
7.5	DNASTAR's Next-Generation Software Solution	91
7.6	Conclusions	94
	References	94

Part Four Emerging Sequencing Technologies 95

8 Real-Time DNA Sequencing 97

Susan H. Hardin

8.1	Whole Genome Analysis	97
8.2	Personalized Medicine and Pharmacogenomics	97
8.3	Biodefense, Forensics, DNA Testing, and Basic Research	98
8.4	Simple and Elegant: Real-Time DNA Sequencing	98
	References	101

9 Direct Sequencing by TEM of Z-Substituted DNA Molecules 103

William K. Thomas and William Glover

9.1	Introduction	103
9.2	Logic of Approach	104
9.3	Identification of Optimal Modified Nucleotides for TEM Visual Resolution of DNA Sequences Independent of Polymerization	106
9.4	TEM Substrates and Visualization	107
9.5	Incorporation of Z-Tagged Nucleotides by Polymerases	108
9.6	Current and New Sequencing Technology	109
9.7	Accuracy	111
9.8	Advantages of ZSG's Proposed DNA Sequencing Technology	111
9.9	Advantages of Significantly Longer Read Lengths	112
9.9.1	<i>De novo</i> Genome Sequencing	112
9.9.2	Transcriptome Analysis	113
9.9.3	Haplotype Analysis	114
	References	115

10 A Single DNA Molecule Barcoding Method with Applications in DNA Mapping and Molecular Haplotyping 117

Ming Xiao and Pui-Yan Kwok

10.1	Introduction	117
10.2	Critical Techniques in the Single DNA Molecule Barcoding Method	118
10.3	Single DNA Molecule Mapping	120
10.3.1	Sequence Motif Maps of Lambda DNA	121
10.3.2	Identification of Several Viral Genomes	123

10.4	Molecular Haplotyping	124
10.4.1	Localization of Polymorphic Alleles Tagged by Single Fluorescent Dye Molecules Along DNA Backbones	125
10.4.2	Direct Haplotype Determination of a Human DNA Sample	127
10.5	Discussion	129
	References	131
11	Optical Sequencing: Acquisition from Mapped Single-Molecule Templates	133
	<i>Shiguo Zhou, Louise Pape, and David C. Schwartz</i>	
11.1	Introduction	133
11.2	The Optical Sequencing Cycle	135
11.2.1	Optical Sequencing Microscope and Reaction Chamber Setup	137
11.2.1.1	Microscope Setup	137
11.2.1.2	Optical Sequencing Reaction Chamber Setup	137
11.2.2	Surface Preparation	137
11.2.3	Genomic DNA Mounting/Overlay	139
11.2.4	Nicking Large Double-Stranded Template DNA Molecules	139
11.2.4.1	Nicking Mounted DNA Template Molecules	139
11.2.4.2	Gapping Nick Sites	139
11.2.5	Optical Sequencing Reactions	140
11.2.5.1	Basic Process	140
11.2.5.2	Choices of DNA Polymerases	140
11.2.5.3	Polymerase-Mediated Incorporations of Multiple Fluorochrome-Labeled Nucleotides	140
11.2.5.4	Washes to Remove Unincorporated Labeled Free Nucleotides and Reduce Background	141
11.2.6	Imaging Fluorescent Nucleotide Additions and Counting Incorporated Fluorochromes	141
11.2.7	Photobleaching	147
11.2.8	Demonstration of Optical Sequencing Cycles	147
11.3	Future of Optical Sequencing	148
	References	149
12	Microchip-Based Sanger Sequencing of DNA	153
	<i>Ryan E. Forster, Christopher P. Fredlake, and Annelise E. Barron</i>	
12.1	Integrated Microfluidic Devices for Genomic Analysis	154
12.2	Improved Polymer Networks for Sanger Sequencing on Microfluidic Devices	156
12.2.1	Poly(<i>N,N</i> -dimethylacrylamide) Networks for DNA Sequencing	156
12.2.2	Hydrophobically Modified Polyacrylamides for DNA Sequencing	159
12.3	Conclusions	160
	References	160

Part Five Next-Generation Sequencing: Truly Integrated Genome Analysis 165

13	Multiplex Sequencing of Paired End Dtags for Transcriptome and Genome Analysis 167
	<i>Chia-Lin Wei and Yijun Ruan</i>
13.1	Introduction 167
13.2	The Development of Paired End Ditag Analysis 168
13.3	GIS-PET for Transcriptome Analysis 170
13.4	ChIP-PET for Whole Genome Mapping of Transcription Factor Binding Sites and Epigenetic Modifications 173
13.5	ChIA-PET for Whole Genome Identification of Long-Range Interactions 175
13.6	Perspective 179
	References 180
14	Paleogenomics Using the 454 Sequencing Platform 183
	<i>M. Thomas P. Gilbert</i>
14.1	Introduction 183
14.2	The DNA Degradation Challenge 184
14.3	The Effects of DNA Degradation on Paleogenomics 185
14.4	Degradation and Sequencing Accuracy 185
14.5	Sample Contamination 189
14.6	Solutions to DNA Damage 191
14.7	Solutions to Contamination 192
14.8	What Groundwork Remains, and What Does the Future Hold? 195
	References 196
15	ChIP-seq: Mapping of Protein–DNA Interactions 201
	<i>Anthony Peter Fejes and Steven J.M. Jones</i>
15.1	Introduction 201
15.2	History 202
15.3	ChIP-seq Method 202
15.4	Sanger Dideoxy-Based Tag Sequencing 203
15.5	Hybridization-Based Tag Sequencing 205
15.6	Application of Sequencing by Synthesis 206
15.7	Medical Applications of ChIP-seq 209
15.8	Challenges 209
15.9	Future Uses of ChIP-seq 211
	References 213
16	MicroRNA Discovery and Expression Profiling using Next-Generation Sequencing 217
	<i>Eugene Berezikov and Edwin Cuppen</i>
16.1	Background on miRNAs 217
16.2	miRNA Identification 218

16.3	Experimental Approach	219
16.3.1	Sample Collection	219
16.3.2	Library Construction	221
16.3.3	Massively Parallel Sequencing	222
16.3.4	Bioinformatic Analysis	223
16.3.4.1	MicroRNA Discovery	223
16.3.4.2	miRNA Expression Profiling	225
16.4	Validation	225
16.5	Outlook	226
	References	226
17	DeepSAGE: Tag-Based Transcriptome Analysis Beyond Microarrays	229
	<i>Kåre L. Nielsen, Annabeth H. Petersen, and Jeppe Emmersen</i>	
17.1	Introduction	229
17.2	DeepSAGE	231
17.3	Data Analysis	235
17.4	Comparing Tag-Based Transcriptome Profiles	235
17.5	Future Perspectives	238
	References	239
18	The New Genomics and Personal Genome Information: Ethical Issues	245
	<i>Jeantine E. Lunshof</i>	
18.1	The New Genomics and Personal Genome Information: Ethical Issues	245
18.2	The New Genomics: What Makes it Special?	245
18.3	Innovation in Ethics: Why do We Need it?	246
18.4	A Proviso: Global Genomics and Local Ethics	247
18.5	Medical Ethics and Hippocratic Confidentiality	247
18.6	Principles of Biomedical Ethics	248
18.7	Clinical Research and Informed Consent	248
18.8	Large-Scale Research Ethics: New Concepts	249
18.9	Personal Genomes	250
18.9.1	What is a Personal Genome and What is New About It?	250
18.9.2	But, Can Making Promises that Cannot be Substantiated be Ever Morally Justifiable?	251
18.10	The Personal Genome Project: Consenting to Disclosure	251
	References	252
	Index	255

Preface

The development of the rapid DNA sequencing method by Fred Sanger and co-workers 30 years ago initiated the process of deciphering genes and eventually entire genomes. The rapidly growing demand for throughput, with the ultimate goal of deciphering the human genome, led to substantial improvements in the technique and was exemplified in automated capillary electrophoresis. Until recently, genome sequencing was performed in large sequencing centers with high automation and many personnel. Even when DNA sequencing reached the industrial scale, it still cost \$10 million and 10 years to generate a draft of the human genome. With the price so high, population-based phenotype–genotype linkage studies were small in scale, and it was hard to translate research into statistically robust conclusions. As a consequence, most presumed associations between diseases and particular genes have not stood up to scientific scrutiny. The commercialization of the first massive parallel pyrosequencing technique in 2004 created the first opportunity for the cost-effective and rapid deciphering of virtually any genome. Shortly thereafter, other vendors entered the market, bringing with them a vision of sequencing the human genome for only \$1000.

This is the topic of this book. We hope to provide the reader with a comprehensive overview of next-generation sequencing (NGS) techniques and highlight their impact on genome research, human health, and the social perception of genetics.

There is no clear definition of next-generation sequencing. There are, however, several features that distinguish NGS platforms from conventional capillary-based sequencing. First, it has the ability to generate millions of sequence reads rather than only 96 at a time. This process allows the sequencing of an entire bacterial genome within hours or of the *Drosophila melanogaster* genome within days instead of months. Furthermore, conventional vector-based cloning, typical in capillary sequencing, became obsolete and was replaced by direct subjecting of fragmented, and usually, amplified DNA for sequencing. Another distinctive feature of NGS are the sequenced products themselves, which are short-length reads between 30 and 400 bp. The limited read length has substantial impact on certain NGS applications, for instance, *de novo* sequencing. The following chapters will present several innovative approaches, which will combine the obvious advantages of NGS, such as

throughput and simplified template preparation, with novel challenging features in terms of short read assembly and large sequencing data storage and processing.

This book arose from the recognition of the need to understand next-generation sequencing techniques and their role in future genome research by the broad scientific community. The chapters have been written by the researchers and inventors who participated in the development and applications of NGS technologies. The first chapter of the book contains an excellent overview on Sanger DNA sequencing, which still remains the gold standard in life sciences. The second and fourth parts of the book describe the commercially available and emerging sequencing platforms, respectively. The third part consists of two chapters highlighting the bottlenecks in the current sequencing: data storage and processing. Once the NGS techniques became available, an unprecedented explosion of applications could be observed. The fifth part of this book provides the reader with the insight into the ever-increasing NGS applications in genome research. Some of these applications are enhancements of existing techniques. Many others are unique to next-generation sequencing marked by its robustness and cost effectiveness, with the prominent example of paleogenomics.

The versatility and robustness of the NGS techniques in studying genes in the context of the entire genome surprised many scientists, including myself. We know that the processes that cause most diseases are not the result of a single genetic failure. Instead, they involve the interaction of hundreds if not thousands of genes. In the past, geneticists have concentrated on genes that have large individual effects when they go wrong, because those effects are so easy to spot. However, combinations of genes that are not individually significant may also be important. It has become evident that next-generation sequencing techniques, together with systems biology approaches, could elucidate the complex dependences of regulatory networks not only on the level of a single cell or tissue but also on the level of the whole organism.

We hope that this book will enrich the understanding of the dramatic changes in genome exploration and its impact not only on research itself but also on many aspects of our life, including healthcare policy, medical diagnostics, and treatment. The best example comes from the field of consumer genomics. Consumer genomics promises to inform people of their risks of developing ailments such as heart disease or cancer; it can even advise its customers how much coffee they can safely drink. This information is retrieved from the correlation of the single nucleotide polymorphism (SNP) pattern of the individual with the SNP haplotype linked to a particular disease. Recent public discussions on the challenges posed by the availability of personal genome information have revealed a new perception of genomic information and its uses. For the first time, a desire to understand the genome has become important and relevant to people outside of the scientific community. In addition to the benefits of having access to genetic information, the ethical and legal risks of making this information available are emerging. The last part of the book introduces the reader to the debate, which will only intensify in the years to come.

In conclusion, I would like to express my sincere gratitude to all of the contributors for their extraordinary effort to present these fascinating technologies and their applications in genome exploration in such a clear and comprehensive way. I also extend my thanks to Professor Hans Lehrach for his constant support.

Berlin, July 2008

Michal Janitz

List of Contributors

Annelise E. Barron

Stanford University
Department of Bioengineering
W300B James H. Clark Center
318 Campus Drive
Stanford, CA 94305
USA

Eugene Berezikov

Hubrecht Institute
Uppsalaalaa 8
3584 CT Utrecht
The Netherlands

Leonard N. Bloksberg

SLIM Search Ltd.
P.O. Box 106-367
Auckland 1143
New Zealand

Edwin Cuppen

Hubrecht Institute
Uppsalaalaa 8
3584 CT Utrecht
The Netherlands

Lei Du

454 Life Sciences
20 Commercial Street
Branford, CT 06405
USA

Tim Durfee

DNASTAR, Inc.
3801 Regent Street
Madison, WI 53705
USA

Jeremy S. Edwards

University of New Mexico Health
Sciences Center
Cancer Research and Treatment Center
Department of Molecular Genetics and
Microbiology
Albuquerque, NM 87131
USA

University of New Mexico
Department of Chemical and Nuclear
Engineering
Albuquerque, NM 87131
USA

Michael Egholm

454 Life Sciences
20 Commercial Street
Branford, CT 06405
USA

Jeppe Emmersen

Aalborg University
Department of Health Science
and Technology
Fredrik Bajers Vej 3B
9000 Aalborg
Denmark

Anthony P. Fejes

Genome Sciences Centre
570 West 7th Avenue, Suite 100
Vancouver, BC
Canada V5Z 4S6

Susan Forrest

University of Queensland
Level 5, Gehrmann Laboratories
Australian Genome Research Facility
St. Lucia, Brisbane, Queensland
Australia

Ryan E. Forster

Northwestern University
Materials Science and Engineering
Department
2220 Campus Drive
Evanston, IL 60208
USA

Christopher P. Fredlake

Northwestern University
Chemical and Biological Engineering
Department
2145 North Sheridan, Tech E136
Evanston, IL 60208
USA

M. Thomas P. Gilbert

University of Copenhagen
Biological Institute
Department of Evolutionary Biology
Universitetsparken 10
2100 Copenhagen
Denmark

William Glover

ZS Genetics
8 Hidden Pond Lane
North Reading, MA 01864
USA

Susan H. Hardin

VisiGen Biotechnologies, Inc.
2575 West Bellfort, Suite 250
Houston, TX 77054
USA

Steven J.M. Jones

Genome Sciences Centre
570 West 7th Avenue, Suite 100
Vancouver, BC
Canada V5Z 4S6

Pui-Yan Kwok

University of California, San Francisco
Cardiovascular Research Institute
San Francisco, CA 94143-0462
USA

University of California, San Francisco
Department of Dermatology
San Francisco, CA 94143-0462
USA

Abizar Lakdawalla

Illumina, Inc.
25861 Industrial Boulevard
Hayward, CA 94545
USA

Jeantine E. Lunshof

VU University Medical Center
EMGO Institute
Section Community Genetics
Van der Boechorststraat 7, MF D424
1007 MB Amsterdam
The Netherlands

Artem E. Men

University of Queensland
 Level 5, Gehrmann Laboratories
 Australian Genome Research Facility
 St. Lucia, Brisbane, Queensland
 Australia

Kåre L. Nielsen

Aalborg University
 Department of Biotechnology,
 Chemistry and Environmental
 Engineering
 Sohngaards-Holms vej 49
 9000 Aalborg
 Denmark

Robert C. Nutter

Applied Biosystems
 850 Lincoln Centre Drive
 Foster City, CA 94404
 USA

Vicki Pandey

Applied Biosystems
 850 Lincoln Centre Drive
 Foster City, CA 94404
 USA

Louise Pape

University of Wisconsin-Madison
 Biotechnology Center
 Departments of Genetics and Chemistry
 Laboratory for Molecular and
 Computational Genomics
 Madison, WI 53706
 USA

Annabeth H. Petersen

Aalborg University
 Department of Biotechnology,
 Chemistry and Environmental
 Engineering
 Sohngaards-Holms vej 49
 9000 Aalborg
 Denmark

Ellen Prediger

Applied Biosystems
 850 Lincoln Centre Drive
 Foster City, CA 94404
 USA

Yijun Ruan

Genome Institute of Singapore
 60 Biopolis Street
 Singapore 138672
 Singapore

David C. Schwartz

University of Wisconsin-Madison
 Biotechnology Center
 Departments of Genetics and Chemistry
 Laboratory for Molecular and
 Computational Genomics
 Madison, WI 53706
 USA

Thomas E. Schwei

DNASTAR, Inc.
 3801 Regent Street
 Madison, WI 53705
 USA

Kirby Siemerling

University of Queensland
 Level 5, Gehrmann Laboratories
 Australian Genome Research Facility
 St. Lucia, Brisbane, Queensland
 Australia

William K. Thomas

Hubbard Center for Genome Studies
 448 Gregg Hall, 35 Colovos Road
 Durham, NH 03824
 USA

Harper VanSteenhouse

Illumina, Inc.
 25861 Industrial Boulevard
 Hayward, CA 94545
 USA

Chia-Lin Wei

Genome Institute of Singapore
60 Biopolis Street
Singapore 138672
Singapore

Peter Wilson

University of Queensland
Level 5, Gehrmann Laboratories
Australian Genome Research Facility
St. Lucia, Brisbane, Queensland
Australia

Ming Xiao

University of California, San Francisco
Cardiovascular Research Institute
San Francisco, CA 94143-0462
USA

Shiguo Zhou

University of Wisconsin-Madison
Biotechnology Center
Departments of Genetics and Chemistry
Laboratory for Molecular and
Computational Genomics
Madison, WI 53706
USA

Part One

Sanger DNA Sequencing

1

Sanger DNA Sequencing

Artem E. Men, Peter Wilson, Kirby Siemering, and Susan Forrest

1.1

The Basics of Sanger Sequencing

From the first genomic landmark of deciphering the phiX174 bacteriophage genome achieved by F. Sanger's group in 1977 (just over a 5000 bases of contiguous DNA) to sequencing several bacterial megabase-sized genomes in the early 1990s by The Institute for Genomic Research (TIGR) team, from publishing by the European Consortium the first eukaryotic genome of budding yeast *Saccharomyces cerevisiae* in 1996 to producing several nearly finished gigabase-sized mammal genomes including our own, Sanger sequencing definitely has come a long and productive way in the past three decades. Sequencing technology has dramatically changed the face of modern biology, providing precise tools for the characterization of biological systems. The field has rapidly moved forward now with the ability to combine phenotypic data with computed DNA sequence and therefore unambiguously link even tiny DNA changes (e.g., single-nucleotide polymorphisms (SNPs)) to biological phenotypes. This allows the development of practical ways for monitoring fundamental life processes driven by nucleic acids in objects that vary from single cells to the most sophisticated multicellular organisms.

"Classical" Sanger sequencing, published in 1977 [1], relies on base-specific chain terminations in four separate reactions ("A", "G", "C", and "T") corresponding to the four different nucleotides in the DNA makeup (Figure 1.1a). In the presence of all four 2'- deoxynucleotide triphosphates (dNTPs), a specific 2',3'-dideoxynucleotide triphosphate (ddNTP) is added to every reaction; for example, ddATP to the "A" reaction and so on. The use of ddNTPs in a sequencing reaction was a very novel approach at the time and gave far superior results compared to the 1975 prototype technique called "plus and minus" method developed by the same team. The extension of a newly synthesized DNA strand terminates every time the corresponding ddNTP is incorporated. As the ddNTP is present in minute amounts, the termination happens rarely and stochastically, resulting in a cocktail of extension

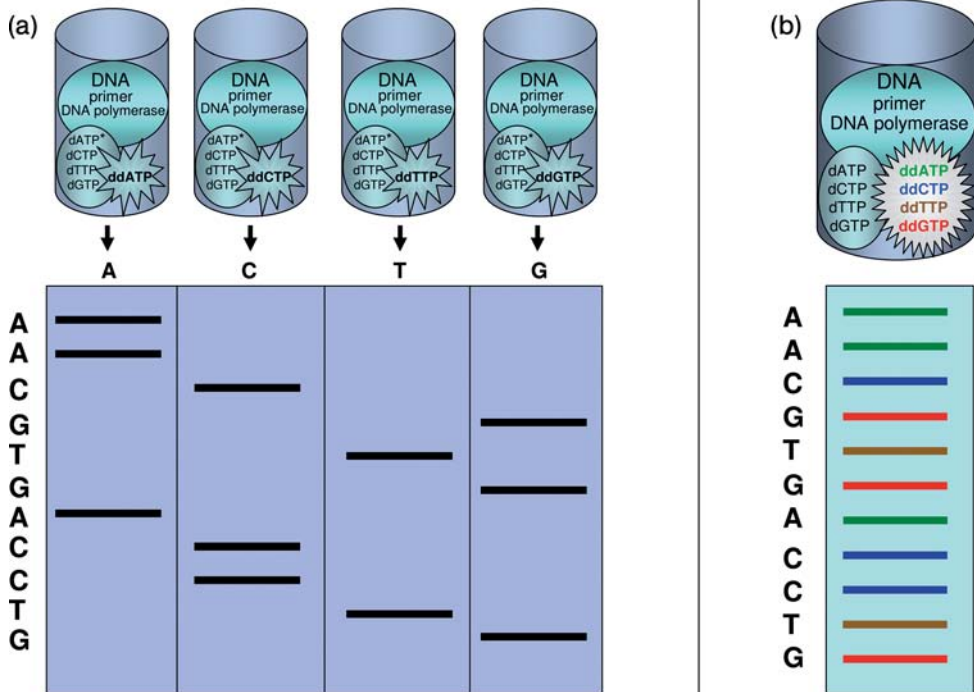


Figure 1.1 Schematic principle of the Sanger sequencing method. (a) Four separate DNA extension reactions are performed, each containing a single-stranded DNA template, primer, DNA polymerase, and all four dNTPs to synthesize new DNA strands. Each reaction is spiked with a corresponding dideoxynucleoside triphosphate (ddATP, ddCTP, ddTTP, or ddGTP). In the presence of dNTPs, one of which is radioactively labeled (in this case, dATP), the newly synthesized DNA strand would extend until the available ddNTP is incorporated, terminating further extension. Radioactive products are then separated through four lanes

of a polyacrylamide gel and scored according to their molecular masses. Deduced DNA sequence is shown on the left. (b) In this case, instead of adding radioactive dATP, all four ddNTPs are labeled with different fluorescent dyes. The extension products are then electrophoretically separated in a single glass capillary filled with a polymer. Similar to the previous example, DNA bands move inside the capillary according to their masses. Fluorophores are excited by the laser at the end of the capillary. The DNA sequence can be interpreted by the color that corresponds to a particular nucleotide.

products where every position of an “N” base would result in a matching product terminated by incorporation of ddNTP at the 3' end.

The second novel aspect of the method was the use of radioactive phosphorus or sulfur isotopes incorporated into the newly synthesized DNA strand through a labeled precursor (dNTP or the sequencing primer), therefore, making every product detectable by radiography. Finally, as each extension reaction results in a very complex

mixture of large radioactive DNA products, probably the most crucial achievement was the development of ways to individually separate and detect these molecules. The innovative use of a polyacrylamide gel (PAG) allowed very precise sizing of termination products by electrophoresis followed by *in situ* autoradiography. Later, the autoradiography was partially replaced by less hazardous techniques such as silver staining of DNA in PAGs.

As innovative as they were 30 years ago, slab PAGs were very slow and laborious and could not be readily applied to interrogating large genomes. The next two major technological breakthroughs took place in (i) 1986 when a Caltech team (led by Leroy Hood) and ABI developed an automated platform using fluorescent detection of termination products [2] separating four-color-labeled termination reactions in a single PAG tube and in (ii) 1990 when the fluorescent detection was combined with electrophoresis through a miniaturized version of PAGs, namely, capillaries [3] (Figure 1.1b). Capillary electrophoresis (CE), by taking advantage of a physically compact DNA separation device coupled with laser-based fragment detection, eventually became compatible with 96- and 384-well DNA plate format making highly parallel automation a feasible reality. Finally, the combination of dideoxy-based termination chemistry, fluorescent labeling, capillary separation, and computer-driven laser detection of DNA fragments has established the four elegant “cornerstones” on which modern building of high-throughput Sanger sequencing stands today.

Nowadays, the CE coupled with the development of appropriate liquid-handling platforms allows Sanger sequencing to achieve a highly automatable stage whereby a stand-alone 96-capillary machine can produce about half a million nucleotides (0.5 Mb) of DNA sequence per day. During the late 1980s, a concept of “highly parallel sequencing” was proposed by the TIGR team led by C. Venter and later successfully applied in human and other large genome projects. Hundreds of capillary machines were placed in especially designed labs fed with plasmid DNA clones around the clock to produce draft Sanger reads (Figure 1.2). The need for large volumes of sequence data resulted in the design of “sequencing factories” that had large arrays of automated machines running in parallel together with automated sample preparation pipelines and producing several million reads a month (Figure 1.3). This enabled larger and larger genome projects to be undertaken, culminating with the human and other billion base-sized genome projects.

Along the way, numerous methods were developed that effectively supported template production for feeding high-throughput sequencing pipelines, such as the whole genome shotgun (WGS) approach of TIGR and Celera, or strategies of subgenome sample pooling of YAC, BAC, and cosmid clones based on physical maps of individual loci and entire chromosomes (this strategy was mainly used by the International Human Genome Project team). Not only did the latter methods help to perform sequencing cheaper and faster but also facilitated immensely the genome assembly stage, where the daunting task of putting together hundreds of thousands of short DNA pieces needed to be performed. Some sophisticated algorithms based on paired end sequencing or using large-mapped DNA constructs, such as finger-printed BACs from physical maps, were developed. Less than 20 years ago,

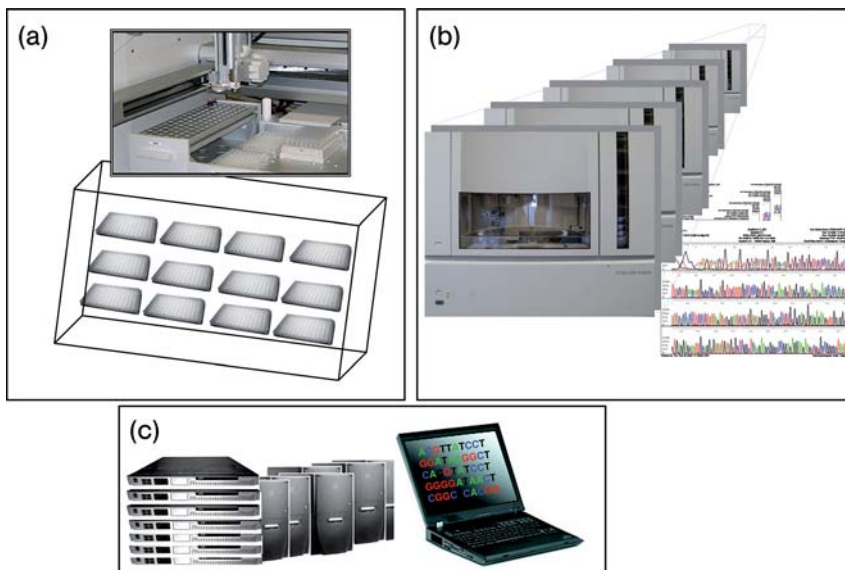


Figure 1.2 Sanger sequencing pipeline. (a) DNA clone preparation usually starts with the isolation of total DNA (e.g., whole genomic DNA from an organism or already fragmented DNA, cDNA, etc.), followed by further fragmentation and cloning into a vector for DNA amplification in bacterial cells. As a result, millions of individual bacterial colonies are produced and individually picked into multiwell plates by liquid-handling robots for isolation of amplified DNA clones. This DNA then goes through a sequencing reaction described in Figure 1.1. (b) Processed

sequenced DNA undergoes capillary electrophoresis where labeled nucleotides (bases) are collected and scanned by the laser producing raw sequencing traces. (c) Raw sequencing information is converted into computer files showing the final sequence and quality of every scanned base. The resultant information is stored on dedicated servers and also is usually submitted into free public databases, such as the GeneBank and Trace Archive.

assembling a 1.8 Mb genome of *Haemophilus influenzae* sequenced by the WGS approach [4] was viewed as a computational nightmare, as it required putting together about 25 000 DNA pieces. Today, a typical next-generation sequencing machine (a plethora of which will be described in the following chapters of this book) can produce 100 Mb in just a few hours with data being swiftly analyzed (at least to a draft stage) by a stand-alone computer.

1.2

Into the Human Genome Project (HGP) and Beyond

The HGP, which commenced in 1990, is a true landmark of the capability of Sanger sequencing. This multinational task that produced a draft sequence published in 2001 [5] was arguably the largest biological project ever undertaken. Now, 7 years later, to fully capitalize on and leverage the data from the Human Genome Project, sequencing technologies need to be taken to much higher levels of output to study

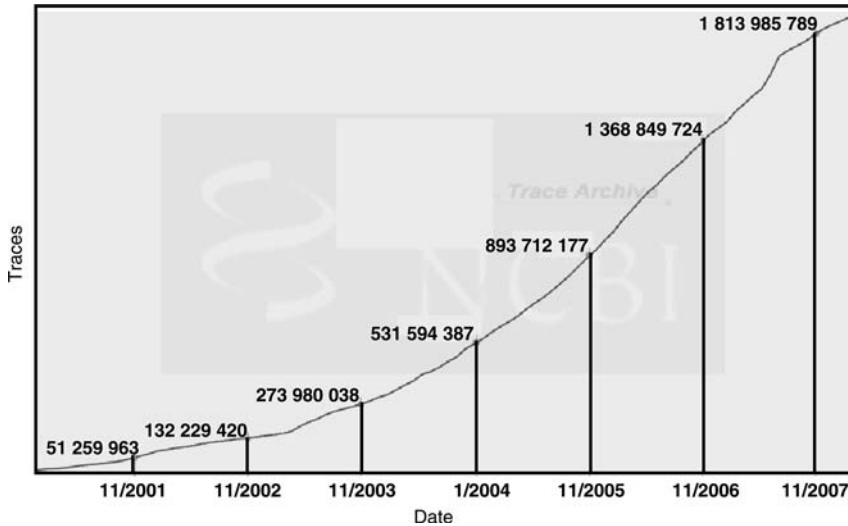


Figure 1.3 Growth of the sequencing information. Number of sequencing traces (reads) submitted to the Trace Archive grew more than 30 times between November 2001 and November 2007. Graph has been modified from reports available at <http://www.ncbi.nlm.nih.gov/Traces/trace.cgi>. Some more statistics of interest: (i) a major genome center produces about 1000 nucleotides per second; (ii) between November 2007 and February 2008, the Trace

Archive received about 200 million trace submissions; (iii) in a single week in February 2008, just the top 10 submissions to the Trace Archive constituted 6 209 892 600 nucleotides; (iv) in 1997, there were 15 finished and published genomes of various sizes and by the end of 2007, there were 710; and (v) there are currently 442 eukaryotic and 965 microbial genome sequencing projects in progress.

multiple genomes cost effectively. Based on the capabilities already available to the medical and other research communities, numerous goals can be envisaged, such as deciphering entire genomes of many individuals, resequencing exons in large cohorts to discover new gene variants, and ultradeep analysis of cellular transcription activities and epigenetic changes that underlie multiple biological phenomena. Opportunities for discovery are virtually endless, from complex diseases to paleogenomics and “museomics” (analysis of ancient DNA), from searching for new organisms in the deep ocean and volcanoes to manipulating valuable traits in livestock and molecular plant breeding. This is where the challenges as well as major opportunities lie in the future.

1.3

Limitations and Future Opportunities

Despite the fact that the Sanger method is still considered by the research community as the “gold standard” for sequencing, it has several limitations. The first is the “biological bias” as the methodology is based on cloning foreign DNA in vectors that

have to be “bacteria friendly” and compatible with the replication machinery of the *E. coli* cells. It has been shown that some parts of chromosomes, such as centromeres and heterochromatic knobs, are practically unclonable. This limitation, in some cases, can be overcome by generating and directly sequencing PCR products but practically it is a very low-throughput and tricky approach. The second challenge is the very restricted ability of Sanger sequencing to handle and analyze allele frequencies. Often, even finding a heterozygous SNP in a PCR product is cumbersome, let alone any bases that are not represented at 1 : 1 ratios.

The third and the most significant burden of the Sanger methodology is the cost. At about \$1 per kilobase, it would cost \$10 000 000 to sequence a 1 Gb genome to 10× coverage! It means that average research laboratories cannot even contemplate sequencing projects that go beyond a megabase scale, thus often totally relying on the large genome centers to get the job done when it comes to sequencing your favorite genome.

Another limitation of Sanger sequencing lies at the genome assembly stage. Although Sanger reads are still the longest on the market, *de novo* assembly of single reads containing repeats is practically impossible without high-resolution physical maps of those regions if a high-quality genome draft is the goal. In regard to the length of a single read, with the current setup of CE separation of dye-tagged extension products, it probably will not reach far beyond 1 kb, despite the development of new fluorophores, with better physical characteristics, and new recombinant polymerases. Nevertheless, further miniaturization of the CE setup or replacing capillaries with chip-based systems with nanochannels that would allow analysis of molecules in the picomolar concentration range, combined with amplification of and signal detection from single template molecules [6], potentially looks like something that will keep Sanger sequencing in the game. In addition, options of combining Sanger outputs with the next-generation reads are quite promising. There of course will be still plenty of low-throughput projects that require only a few reads to be performed for a particular task, for which Sanger sequencing undoubtedly is an excellent and mature technology and will remain the gold standard for quite some time.

1.4

Bioinformatics Holds the Key

In the past 5 years, about a dozen genomes larger than a billion nucleotides in size were sequenced and assembled to various finished stages. There are 905 eukaryotic genomes currently in production as of February 2008 (<http://www.genomesonline.org/gold.cgi>). Most importantly, every new bit of data is being immediately made available to the general research community through databases such as the Trace Archive (<http://0-www.ncbi.nlm.nih.gov.catalog.llu.edu/Traces/trace.cgi> and <http://www.tracearchive.ntu.ac.uk>) (Figure 1.3). This enormous terabyte-sized data flow generates previously unseen possibilities for computer-based analysis and boosts fields such as comparative and population genomics to new levels of biological

discoveries via *in silico* data manipulation. The importance of the role of the bioinformatician as a major player in modern biology cannot be understated, and it will only grow with the advent of next-generation sequencers and sequencing pipelines. The larger genome projects already undertaken with Sanger sequencing have required the development of many analytical algorithms and quality assessment tools. With the significant growth in the output of DNA sequence information from the Sanger method to the next-generation DNA sequencers comes a concomitant rise in the amount of sequence information to be checked, assembled, and interpreted.

1.5 Where to Next?

A few questions obviously stand out. How can we move from just the ability to sequence a genome of a chosen individual to practical solutions that can be used in population studies or personalized medicine? How to use DNA-based information in routine medical checkups and for personalized drug prescription based on prediction of potential diseases? How one can access and explore genetic diversity in a given population, whether it is a study of diversity in birds or a search for new drought-resistant traits in agricultural crops?

The impact of genome sequencing on everyday life is getting more and more obvious. Making it affordable is the next big challenge. Many methods aimed at decreasing the cost of individual sequences are being developed very rapidly, such as genome partitioning through filtering or hybridization processes that reduce the complexity of the DNA sample to its most informative fraction, say a set of particular exons. From early experiments that involved filtration for nonrepetitive DNA via DNA reassociation followed by the sequencing of the nonrepetitive fraction to recently published array-based capturing of a large number of exons [7–9], the “genome reduction” concept seems to hold one of the keys to cheaper sequencing, as it strips down the complexity of a given genome to its gene-coding essence (almost two orders of magnitude) making it more readily accessible to sequencing analysis. The hype about developing new, cheaper ways for deciphering individual genomes was certainly boosted by several prizes offered for developing new platforms, with the paramount goal being the “\$1000-genome” mark [10]. The book you hold in your hands is dedicated to the most recent ideas of how this goal might be achieved. It presents a number of totally new, exciting approaches taken forward in just a last few years that have already contributed immensely to the field of sequencing production in general and cost-efficiency in particular. Although not in the scope of this introductory chapter, it is truly worth mentioning that up to a 500 megabases a day is the average productive capacity of a current next-generation sequencing platform that is at least a thousand times more efficient than the standard 96-capillary machines used for the HGP.

The growing power of genome sequencing from the ability to sequence a bacteriophage in the late 1970s to a bacterial genome and then finally to human genome suggests that the sequencing capacity increases by about three orders of

magnitude every decade. It is hard to predict which method(s) will dominate the sequencing market in the next decade, just as it was hard to predict 30 years ago whether Sanger's or Maxam–Gilbert's method would become a major player. Both methods were highly praised in their initial phase and secured Nobel prizes for both team leaders. At that time, probably nobody would have been able to predict that the Sanger sequencing would take preference over the Maxam–Gilbert technology as a method of choice, largely thanks to subsequent development and application of shotgun cloning, PCR, and automation. In any case, only time will tell whether the next champion in DNA sequencing production will be a highly parallel data acquisition from hundreds of millions of short DNA fragments captured in oil PCR “nanoreactors,” or attached to a solid surface (see following chapters) or individual analysis of unlabeled and unamplified single nucleic acids with data collection in real time, based on their physical changes detected by Raman or other spectral methods [11].

Rough extrapolation suggests that, with the current progress of technology, in 2020, we will be able to completely sequence a million individuals or produce a hundred million “exome” data sets (assuming one set being 20 000 gene exons of 1.5 kb each). Quite an impressive number, but regardless of cost it still is only about 1% of the planet's population. Nevertheless, the recent online announcement from U.S. and Chinese scientists of a very ambitious plan of sequencing up to 2000 human genomes in the coming 3 years (http://www.insequence.com/issues/2_4/features/144575-1.html) has set the bar much higher for every aspect of the technology, from the accumulation of reads to the analysis and storage of terabytes of data. It is an exciting time in genome biology, and the combination of existing sequencing methods such as Sanger and the numerous next-generation sequencing tools will result in a wealth of data ready for mining by intrepid bioinformaticians and then given back to scientists, doctors, criminologists, and farmers.

References

- 1 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.
- 2 Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S. and Hood, L. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674–679.
- 3 Swerdlow, H. and Gesteland, R. (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research*, **18**, 1415–1419.
- 4 Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., McKenney, K., Sutton, G., FitzHugh, W., Fields, C. and Venter, J. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **268**, 496–498.
- 5 International Human Genome Sequencing Consortium . (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- 6 Xiong, Q. and Cheng, J. (2007) Chip capillary electrophoresis and total genetic analysis systems, in *New High Throughput*

- Technologies for DNA Sequencing and Genomics* (ed. K.R. Mitchelson), Elsevier, Amsterdam.
- 7 Albert, T., Molla, M., Muzny, D., Nazareth, L., Wheeler, D., Song, X., Richmond, T., Middle, C., Rodesch, M., Packard, C., Weinstock, G. and Gibbs, R. (2007) Direct selection of human genomic loci by microarray hybridization. *Nature Methods*, **4**, 903–905.
 - 8 Okou, D., Stinberg, K., Middle, C., Cutler, D., Albert, T. and Zwick, M. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nature Methods*, **4**, 907–909.
 - 9 Porreca, G., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S., LeProust, E., Peck, B., Emig, C., Dahl, F., Gao, Y., Church, G. and Shendure, J. (2007) Multiplex amplification of large sets of human exons. *Nature Methods*, **4**, 931–936.
 - 10 Bennett, S., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005) Toward the \$1000 human genome. *Pharmacogenomics*, **6**, 373–382.
 - 11 Bailo, E. and Deckert, V. (2008) Tip-enhanced Raman spectroscopy of single RNA strands: towards a novel direct-sequencing method. *Angewandte Chemie*, **4**, 1658–1661.

Part Two

Next-Generation Sequencing: Toward Personalized Medicine

2

Illumina Genome Analyzer II System

Abizar Lakdawalla and Harper VanSteenhouse

With the ability to sequence more than 60 000 000 DNA fragments simultaneously, the first generation of the Illumina Genome Analyzer had revolutionized the ability of labs to generate large volumes of sequence data in just a week at an extremely low cost; data that previously required the resources of a genome center, the efforts of many scientists over many months, and an expenditure of millions of dollars [1, 2]. The diverse range of applications facilitated by the Genome Analyzer's massively parallel sequencing technology, simple workflow, and a 100× decrease in cost brings us significantly closer to understanding the links between genotype and phenotype and in establishing the molecular basis of many diseases [3].

The current version, the Genome Analyzer II, generates gigabases of high-quality data per day with an uncomplicated process that requires just one operator and less than 6 h of hands-on time (Figure 2.1). Specifications for the Genome Analyzer II are summarized in Table 2.1.

The sequencing-by-synthesis process used in the Genome Analyzer II consists of the following three stages:

1. DNA library preparation with Illumina sample prep kits;
2. automated generation of clonal clusters on the cluster station; and
3. Sequencing of $>5 \times 10^7$ clonal clusters on the Genome Analyzer.

2.1

Library Preparation

DNA fragmentation: DNA from whole genomes, metagenomes, or long PCR is fragmented by nebulization or sonication. DNA fragments may also be isolated from targeted regions of the genome [4–6], cDNA, enzyme-digested DNA, bisulfite-treated DNA, or from DNA coimmunoprecipitates [7]. The library preparation

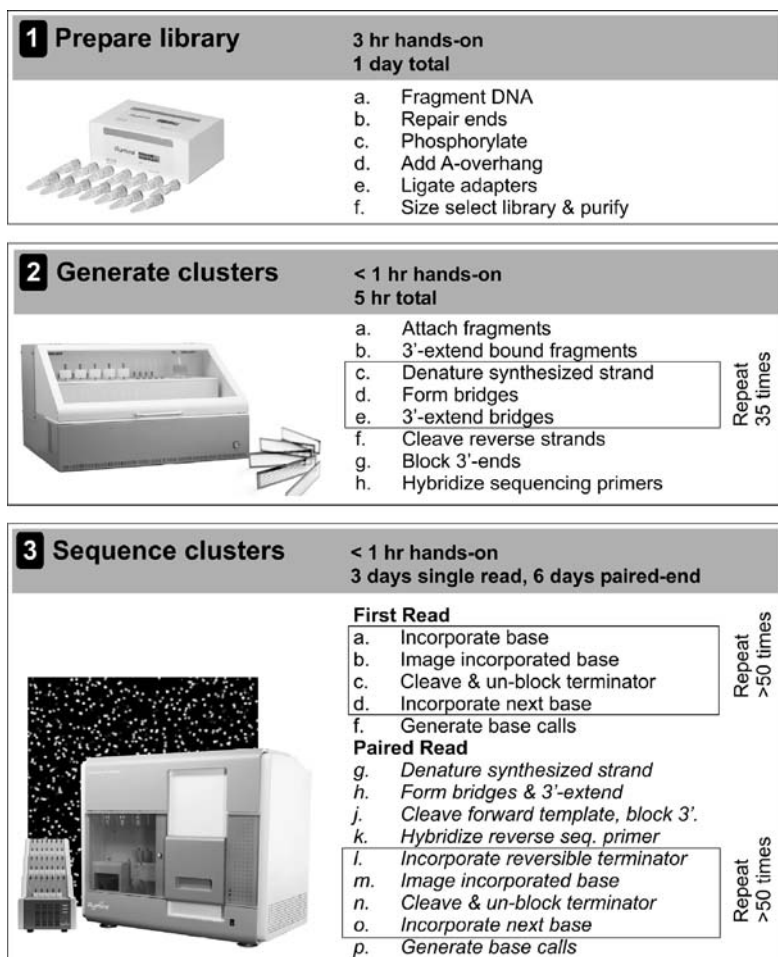


Figure 2.1 Genome Analyzer II sequencing system and workflow. (1) Sample prep kits for creating libraries, (2) automated Cluster Station for generating clonal clusters in a flow cell, and (3) sequencing in forward and reverse directions on the Genome Analyzer with the integrated paired end module.

procedure requires approximately 100 ng of genomic DNA and takes ≈ 1 day (≈ 3 h hands-on). Libraries generated from chromatin immunoprecipitates require about 10 ng of DNA.

End repair and ligation: DNA fragments are blunt ended by polymerase and exonuclease activity (Figure 2.2a) followed by phosphorylation (Figure 2.2b) and addition of an adenosine overhang (Figure 2.2c). Illumina-sequencing adapters are then ligated to the fragments (Figure 2.2d). A fragment library of the desired molecular weight is isolated by gel electrophoresis and PCR selection.

Table 2.1 Genome Analyzer II specifications (as of July 2008).^a

System components

- Library prep kits for genomic DNA, gene expression analysis (digital gene expression), small RNA, and chromatin immunoprecipitates (ChIP-seq)
- Cluster station: Fluidics system for an automated formation of clonal clusters. Capacity: 1 flow cell (17 mm × 66 mm with eight independent sealed channels), 18 reagents + 1 waste, factory and user-defined protocols. Cluster generation kits, for single and paired end sequencing, include flow cells and reagents
- Genome Analyzer II: Imaging and fluidics system for an automated sequencing of clonal clusters. Bench-top system, 1 flow cell capacity, 3 laser illumination system. Utilizes universal sequencing kits for all applications
- Paired end module: Attachment for the Genome Analyzer to enable paired end sequencing
- IPAR analysis system: Data analysis computer with 32 GB RAM, 8 kernels, 9TB RAID, with real-time data QC, and automatic data transfer

Throughput

- Sequence output of ≈1500 megabases/day, ≈3700 megabases/single-read run, ≈7500 megabases/paired end run
- Run time of 2.5 d for a single read (50 bases), 5.5 d for paired ends (2 × 50 bases)
- Cluster prep: Less than 1 h hands-on, 5 h total, up to eight samples per flow cell, >50 million productive clonal clusters per flow cell
- Library prep: 3 h hands-on, 6 h total for genomic DNA libraries

Accuracy

- Accuracy of ≈99.999% at 3× or greater coverage (with 36 bp paired end reads in *E. coli* K12)
- Raw accuracy of ≈98.5 with ≈90% of reads with zero errors

Sample amount

- ≈2 pM DNA fragments per flow cell, ≈100 ng of genomic DNA for library preparation

^aSpecifications are subject to change, please visit illumina.com for most up-to-date information.

2.2 Cluster Creation

Template molecules are attached to a flow cell for generation of clonal clusters and for sequencing cycles. The flow cell is a silica slide with eight lengthwise lanes. A separate sample library can be added to each lane, enabling eight separate sequencing runs per flow cell. The sealed flow cell minimizes the risk of contamination and handling errors. The clonal clusters are generated on an Illumina Cluster Station (Figure 2.1) in ≈5 h (<1 h hands-on). The process does not require clean rooms, robotics, or additional hardware.

The DNA fragment library is diluted to ≈2 pM (fragments), denatured and introduced into one to eight channels of the flow cell. Forward and reverse strands are captured by a lawn of primers covalently attached to the flow cell surface

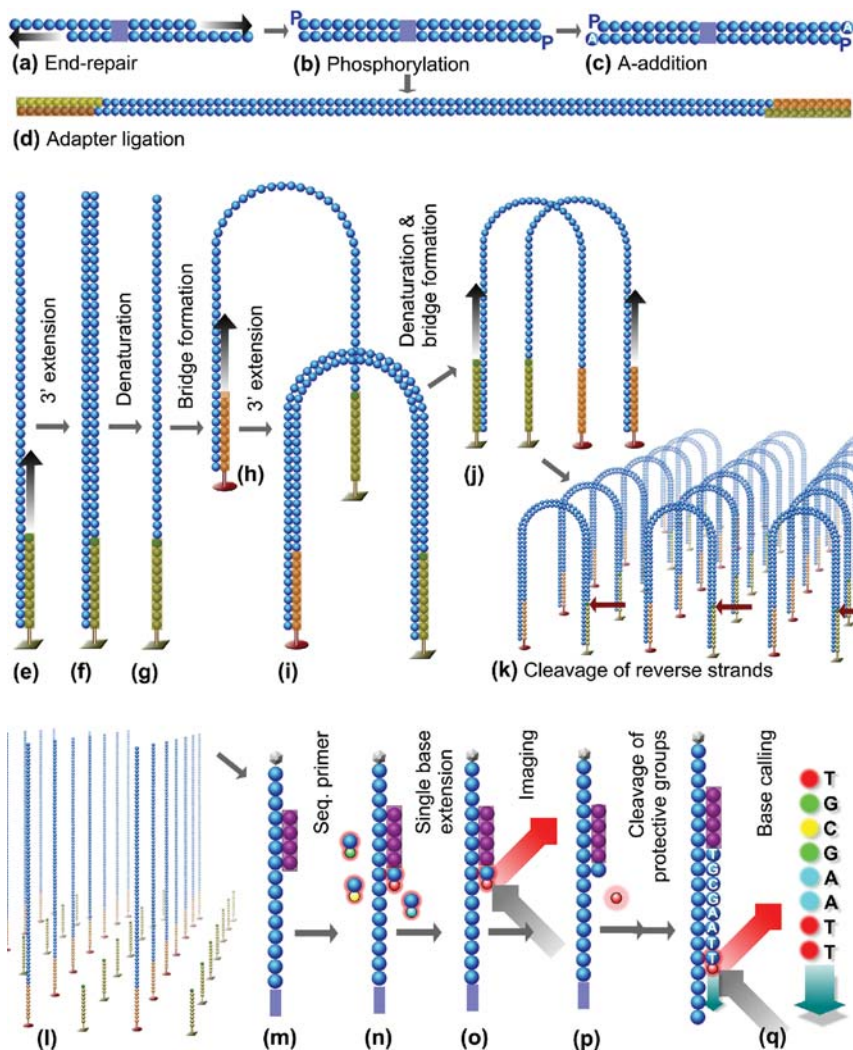


Figure 2.2 The sequencing-by-synthesis process. *Library preparation:* DNA fragments are blunt ended (a), phosphorylated (b), an A-overhang is added (c), and then ligated to adapters (d). *Cluster formation:* Denatured DNA is bound to primers on the flow cell surface (e), extended (f), and then denatured (g). The ssDNA forms a bridge to adjoining lawn primers (h), is extended again to create a double-stranded bridge (i), denatured and hybridized to form new ssDNA bridges (j). The process is repeated 35

times to give discrete clusters with ≈ 2000 molecules (k). Reverse strands are cleaved (arrow, k), 3' ends are blocked and the strands are hybridized to sequencing primers (m). *Sequencing:* DNA polymerase incorporates a reversible terminator (n), the fluorescent base is imaged (o), the fluorescent tag is then cleaved off, and the terminator is unblocked (p). The process is repeated for ≥ 50 cycles to determine the sequence (q).

(Figure 2.2e). Templates bound to the primers are 3'-extended (Figure 2.2f) producing covalently attached discrete single molecules (Figure 2.2g). Free ends of the bound templates hybridize to adjacent lawn primers to form U-shaped bridges (Figure 2.2h). The DNA bridge is copied from the primer to create a double-stranded DNA (dsDNA) bridge (Figure 2.2i). The resulting dsDNA is denatured, hybridized to lawn primers to form new bridges, and extended again (Figure 2.2j). This process of isothermal bridge amplification is repeated 35 times to create a dense cluster of ≈ 2000 molecules (Figure 2.2k). The reverse strands in the cluster are removed by cleavage at the reverse strand-specific lawn primers (Figure 2.2k). The clusters are denatured and free 3'-OH ends are blocked to prevent nonspecific priming (Figure 2.2l), and sequencing primers are hybridized to the free ends of the DNA templates (Figure 2.2m).

The flow cell with the synthesized clonal clusters is then transferred to the Genome Analyzer instrument for sequencing.

2.3

Sequencing

Sequencing on the Genome Analyzer II consists of three stages:

- incorporation of fluorescent reversible terminators by DNA polymerase (Figure 2.2n);
- imaging of the fluorescently labeled clusters (Figure 2.2o); and cleavage and unblocking of the incorporated nucleotide before the next addition cycle (Figure 2.2p).

The single-base extension cycle is repeated for ≥ 50 cycles (Figure 2.2q) to generate the DNA sequence of the $>6 \times 10^7$ clusters simultaneously.

The DNA polymerase incorporates a specific reversible terminator in the presence of all four bases with high accuracy as the four nucleotides compete for the polymerase. Since each cycle is terminated and read before moving to the next cycle, homopolymer sequences are determined precisely. The high levels of G–C bias observed with hybridization and ligation approaches are not observed in sequencing-by-synthesis methods. Sequencing-by-synthesis is also compatible with DNA sequences of reduced complexity, such as bisulfite-treated DNA.

2.4

Paired End Reads

Paired end reads produce informative data for consistent alignment of sequence reads. The increased data output and length coverage with paired end reads facilitate *de novo* assembly, detection of indels, inversions, and other genome rearrangements, and directional whole transcriptome sequencing.

On completion of the sequencing of the forward strand, the newly synthesized partial strand is removed by denaturation, the 3' ends are unblocked, and double-

stranded DNA clusters are regenerated by bridge amplification. The forward strands are removed by cleavage at the base of the forward strands leaving only the newly synthesized reverse strands attached to the flow cell. The reverse strands are then sequenced as before to produce paired end sequence data.

2.5

Data Analysis

The Genome Analyzer comes with a comprehensive suite of software, the Sequencing Control Studio (SCS), the Genome Analyzer Pipeline, and the GenomeStudio analysis software. The SCS manages the Genome Analyzer instrument, provides detailed first-base incorporation statistics, and real-time performance data throughout the run (Figure 2.3). The Genome Analyzer Pipeline

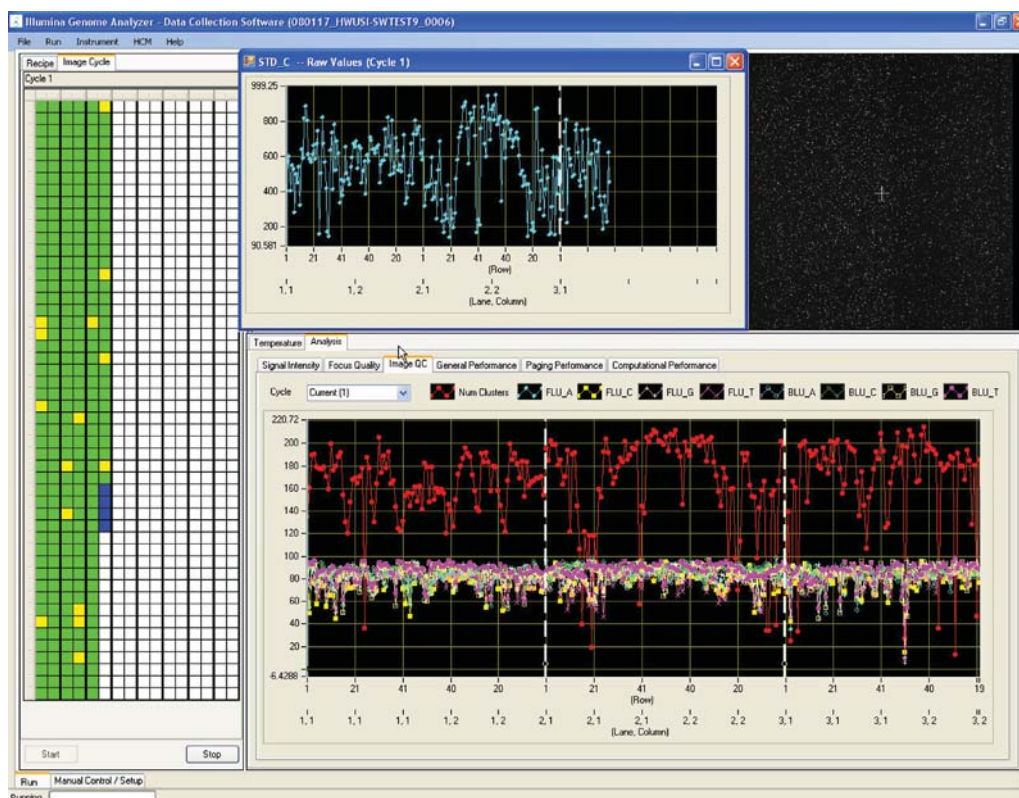


Figure 2.3 The SCS Run Browser module. Real-time performance statistics – number of clusters, fluorescent intensities, and so on – are monitored to provide a snapshot of performance during the run. Activity for all sectors of a flow cell is indicated in the form of heat maps for a comprehensive overview.

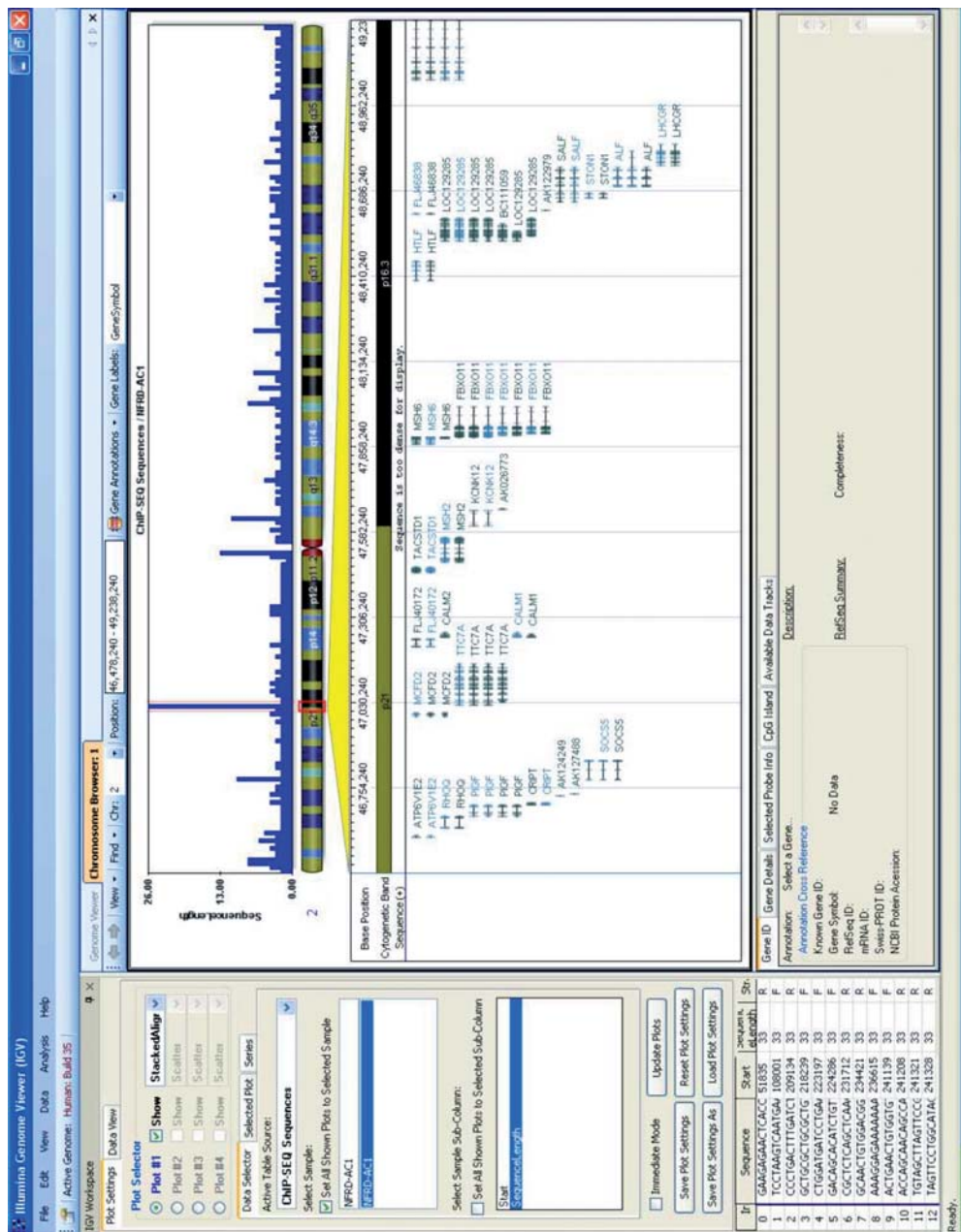
software consists of a series of modules that work together to enable a seamless data analysis workflow. The Pipeline software performs the image analysis, identification and alignment of clusters, assignment of quality scores, filtering, base calling, and alignment. Automated image calibration ensures that a maximum number of clusters are used to generate data. Accurate cluster intensity scoring algorithms deliver high-quality reads. Quality-calibrated base calls allow high-confidence detection of true polymorphisms. Intelligent paired end logic provides a tool for identifying structural variants and for sequencing repetitive regions. The raw and aligned data can be exported into a number of analysis tools. The Illumina GenomeStudio provides an easy-to-use software suite for the analysis of protein–nucleic acid interactions, SNP discovery, and other emerging applications (Figure 2.4).

A partial list of software tools for the Genome Analyzer is given below:

- Sequence assembly tools:
 - Euler-SR (Pevzner, Chaisson; UC San Diego): Genomic assembly;
 - MAQ (Mapping and Assembly with Quality; Heng Li, Sanger Centre): Alignment and polymorphism detection;
 - MUMmerGPU: Sequence alignment with Graphics Processing units [8];
 - SHARCGS (Max Planck Institute for Molecular Genetics): Short-read assembly algorithm [9];
 - SSAKE (British Columbia Cancer Agency, Genome Sciences Centre): Assembly of short reads [10];
 - Velvet (Zerbino and Birney, EMBL-EBI): *De novo* assembly of short reads.
- Genome viewers:
 - Gbrowse: Genomic Browsing Generic Model Organism Database Project [10];
 - Staden Tools (GAP4) (Cambridge University): Alignment and visualization tools for small data sets [11];
 - UCSC Browser: Genome browsing and comprehensive annotation.
- Tag-based sequencing and digital gene expression (DGE):
 - ChIP-Seq Peak Finder (Cal Tech): Mapping and counting of DNA fragments to the genome isolated by Chromatin immunoprecipitation [12];
 - Digital Gene Expression Comparative Count Display (NIH): Gene expression analysis by sequencing of cDNA tags;
 - SAGE DGED Tool (Cancer Genome Anatomy Project, NCI, NIH): Gene expression analysis by sequencing of cDNA tags.

2.6 Applications

In this section, a brief overview of the primary applications of the Genome Analyzer for whole genome and targeted genome sequencing, epigenomics, transcriptome profiling, protein–nucleic acid interactions, and multiplexing samples is provided.



2.6.1

Genome Sequencing Applications

Genomes being sequenced with the Genome Analyzer range from *de novo* assembly or resequencing of *Staphylococcus*, *Helicobacter*, *Escherichia*, *Pseudomonas*, *Mycobacteria*, microbial communities, *Arabidopsis*, *Caenorhabditis* to human genome. Targeted regions of the genome such as the immunoglobulin switching regions, regions identified by association studies, or all the coding exons can be isolated by a series of long PCR reactions or by utilizing the large number of oligonucleotides synthesized on arrays to capture and amplify exonic regions [4–6].

The minimal sample requirements (≤ 2 pM fragments per run) enable the sequencing of laser captured cells, limited archival tissues, embryoid bodies, small model systems, and difficult-to-cultivate organisms such as *Microsporidia*.

The large volumes of genomic sequencing data generated are being used to discover genomic variation in the human genome, such as copy number variations (CNVs) (Figure 2.5), chromosomal rearrangements (deletions, insertions, and translocations), and single-nucleotide variations, to discover the genetic cause of drug resistance in pathogens, to understand coevolution of endosymbionts, and so on [13, 14].

2.6.2

Epigenomics

Variation in DNA methylation across the whole genome is being determined by sequencing of fragments generated by methylation restriction digests, purification of methylated fragments by antibody affinity, or by bisulfite sequencing [7]. Active DNA methylation by trapping methyltransferases to DNA labeled *in vivo* with azanucleotides can also be used to determine the pattern of active methylation. In addition, patterns of histone modifications at promoters, insulators, enhancers, and transcribed regions have been established at very high resolution in the human genome (Figure 2.6) and linked to gene activation and repression [15].

2.6.3

Transcriptome Analysis

An exciting area of functional genomics is the characterization of ALL transcriptional activity, coding and noncoding, without any prior assumptions. Sequencing of cDNA



Figure 2.4 Illumina GenomeStudio data analysis and genome viewer. Developed specifically for tag-based sequencing applications on the Genome Analyzer, GenomeStudio presents genome-scale data in the form of user-friendly graphical views with the ability to zoom in from the level of chromosomes to a single base. The y-axis of the graph shows the number of reads mapping to base positions (x-axis) for a library prepared from a chromatin immunoprecipitation experiment.

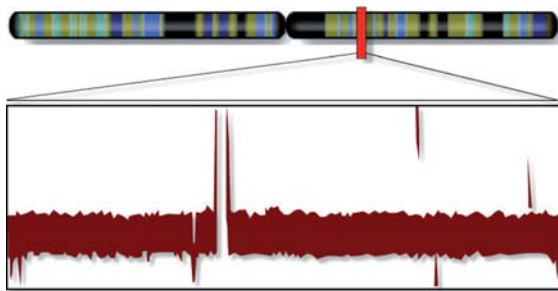


Figure 2.5 Copy number variation in the human X-chromosome. Paired end sequence data were generated from an isolated Caucasian X-chromosome according to standard Genome Analyzer protocols. CNVs are indicated by a substantial change in the number of reads at base locations. The graph is redrawn from representative data.

libraries generated from random primed extensions of RNA fractions (nuclear, cytoplasmic, poly-A, capped, small RNA, and large RNA) provides a high-resolution map of the transcriptome for the detection of uncharacterized RNA species, discovery of novel tissue-specific isoforms and cSNPs, unambiguous assignment of 5' exons and UTRs (Figure 2.7) [7, 16], and the characterization of previously undetectable microRNA species [17].

The massive throughput of >60 M reads offers an exquisite sensitivity and an unprecedented dynamic range for quantitative applications. A single copy of RNA in

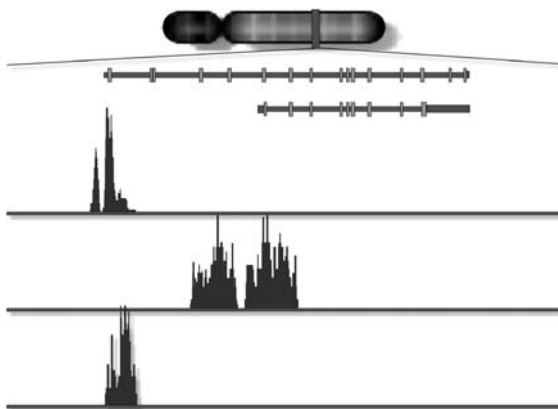


Figure 2.6 ChIP-seq analysis of histone modifications on human Chr 22. Sequences generated from chromatin immunoprecipitates with antibodies specific to modified histones (K4, K36, and K79) align specifically to promoter regions. Redrawn from representative data.

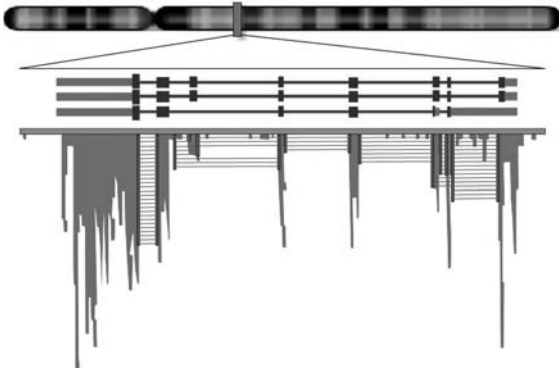


Figure 2.7 Discovery of alternative transcript splice forms. Alignment of sequence fragments from a random-primed library of poly-A RNA shows a high number of reads aligning to exonic regions. A proportion of the reads span introns. The relative concentration of each splice form can be deduced from the number of reads that are unique to that isoform [7].

a cell can be detected in $<1\ \mu\text{g}$ of total RNA by digital gene expression methods (sequence tags created by restriction enzyme digestion of cDNA). The number of tags sequenced gives an absolute count (from one to a few million) of all RNA in the sample (Table 2.2) with accuracy comparable to qPCR. As DGE does not require any a priori sequence information, it can be used for species with inadequate reference genomes, and as tag databases are revised, existing data can be readily reanalyzed.

Table 2.2 Digital gene expression analysis of poly-ARNA from human brain and universal human control RNA.

Symbol	Name	Tag	Count in human brain	Count in universal human control
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	GATCATGAGCAATGCCTCCT	10 202	21 582
HBZ	Hemoglobin, Zeta	GATCTCCACGCAGGCCCGACA	0	4443
IGJ	Immunoglobulin J polypeptide	GATCCTACAGAAGTGGAGCT	0	113
PRDX2	Peroxisredoxin 2	GATCACTGTTAATGATTTC	162	163
RPL23	Ribosomal protein L 29	GATCAAACCAAGGCCAGGC	118	861
STMN4	Stathmin-like 4	GATCATCATGACCAATGAAA	73	0

Count = templates per million.

2.6.4

Protein–Nucleic Acid Interactions

Sequencing of DNA bound to immunoprecipitates provides a highly effective method to ascertain whole genome patterns of occupancy by nucleosomes, polymerases, and transcription factors [7, 18–23]. In chromatin immunoprecipitation sequencing (ChIP-seq) protocols, sample libraries are generated by fixing chromatin with formaldehyde, fragmenting the chromatin, isolating specific fragments with an antibody or an affinity binder, extracting the DNA, and finally preparing libraries. The libraries are sequenced and the fragments are mapped back to the genome to locate regions of interest. A modification of the ChIP-seq method can be applied to elucidate RNA–protein interactions, for example, in RNAi complexes.

2.6.5

Multiplexing

Many of the applications summarized above are amenable to sample multiplexing. The ability to multiplex samples increases the efficiency of the system by enabling the analysis of tens to hundreds of samples simultaneously. Indexing individual samples requires the inclusion of an indexing tag to the adapter oligonucleotides that is unique to the sample. Samples would be prepared separately, ligated to sample-specific adapters (index sequence + universal adapter), and then combined together for cluster formation and sequencing. For example, a four-base index code that would allow 256 individual codes (4^4 unique combinations) would therefore enable 256-plex sequencing.

2.7

Conclusions

The Genome Analyzer has enabled many scientists to perform studies that are revolutionizing our understanding of the molecular biology of disease.

- Studying the epigenomic and genomic variations associated with and causal to many diseases.
- Creating a detailed catalog of all transcripts, and how they change with disease.
- Discoveries of how the replication and transcription machinery, transcription factors, and gene expression modulators interact with the genome.

The first generation of the Genome Analyzer has succeeded in transforming the functional genomics landscape by providing a $100\times$ increase in sequencing throughput with a concomitant decrease in the overall cost of sequencing. The new Genome Analyzer II offers even greater system robustness, increased throughput, and the highest quality data of any next-generation sequencing system available today while still maintaining the easiest workflow and the lowest operating costs.

References

- 1 Bentley, D.R. (2006) Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, **16** (6), 545–552.
- 2 Mardis, E.R. (2006) Anticipating the 1,000 dollar genome. *Genome Biology*, **7** (7), 112.
- 3 Blow, N. (2007) The personal side of genomics. *Nature*, **449**, 627–630.
- 4 Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., Leproust, E.M., Peck, B.J. *et al.* (2007) Multiplex amplification of large sets of human exons. *Nature Methods*, **4**, 931–936.
- 5 Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J. and McCombie, W.R. (2007) Genome-wide *in situ* exon capture for selective resequencing. *Nature Genetics*, **39**, 1522–1527.
- 6 Olson, M. (2007) Enrichment of super-sized resequencing targets from the human genome. *Nature Methods*, **4**, 891–892.
- 7 Wold, B. and Myers, R.M. (2008) Sequence census methods for functional genomics. *Nature Methods*, **5** (1), 19–21.
- 8 Schatz, M.C., Trapnell, C., Delcher, A.L. and Varshney, A. (2007) High-throughput sequence alignment using Graphics Processing Units. *BMC Bioinformatics*, **8** (1), 474.
- 9 Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Research*, **17**, 1697–1706.
- 10 Warren, R.L., Sutton, G.G., Jones, S.J. and Holt, R.A. (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **23** (4), 500–501.
- 11 Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. and Lewis, S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Research*, **12** (10), 1599–1610.
- 12 Staden, R., Judge, D.P. and Bonfield, J.K. (2003) Managing sequencing projects in the GAP4 environment, in *Introduction to Bioinformatics: A Theoretical and Practical Approach* (eds S.A. Krawetz and D.D. Womble), Human Press Inc., Totawa, NJ, p. 07512.
- 13 Topol, E.J. and Frazer, K.A. (2007) The resequencing imperative. *Nature Genetics*, **39** (4), 439–440.
- 14 McCutcheon, J.P. and Moran, N.A. (2007) Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, **104** (49), 19392–19397.
- 15 Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
- 16 Wakaguri, H., Yamashita, R., Suzuki, S., Sugano, S. and Nakai, K. (2008) DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Research*, **36**, D97–D101, 10.1093/nar/gkm901.
- 17 Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., Holoch, D., Lim, C. and Tuschl, T. (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, **44** (1), 3–12.
- 18 Mardis, E.R. (2007) ChIP-seq: welcome to the new frontier. *Nature Methods*, **4** (8), 651–657.
- 19 Fields, S. (2007) Site-seeing by sequencing. *Science*, **316**, 1441–1442.
- 20 Schmid, C.D. and Bucher, P. (2007) ChIP-seq data reveal nucleosome architecture of human promoters. *Cell*, **131** (5), 831–832.
- 21 Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W. *et al.* (2007)

Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.

Nature, **448**, 553–560.

- 22 Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using

chromatin immunoprecipitation and massively parallel sequencing.

Nature Methods, **4**, 651–657.

- 23 Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of *in vivo* protein DNA interactions. *Science*, **316** (5830), 1497–1502.

3

Applied Biosystems SOLiD™ System: Ligation-Based Sequencing

Vicki Pandey, Robert C. Nutter, and Ellen Prediger

3.1

Introduction

The development of massively parallel sequencing systems is dramatically altering the way scientists conduct their research. As the cost per base of DNA sequence generated continues to decrease, the number of bases sequenced per run has steadily increased. The Applied Biosystems next-generation SOLiD™ System sequencing technology is based on sequential ligation of dye-labeled oligonucleotides, enabling massively parallel sequencing of clonally amplified DNA fragments. Features inherent to this system, such as mate-paired analysis and two-base encoding, enable studies of complex genomes by providing a greater degree of accuracy. In addition, this system contains the necessary computing power and software to perform base calling on a large scale without the need for additional computing hardware. The throughput and scalability of the SOLiD System holds great promise for large-scale resequencing, digital gene expression, and hypothesis-free chromatin immunoprecipitation (ChIP) and methylation studies. The SOLiD System includes the SOLiD Analyzer, which is comprised of instrument components, computer system, and SOLiD Software Suite (Figure 3.1).

3.2

Overview of the SOLiD™ System

The performance of the SOLiD System results from three innovative, leading-edge features: high-fidelity ligase chemistry, primer reset function, and two-base encoding. This combination produces a highly robust system with the accuracy necessary to support high-throughput variation detection. Several additional features – probe specificity, highly efficient chemical cleavage of the ligated probes, and dye-labeled sequencing products – further differentiate the technology of the SOLiD System from that of its competitors, rendering it unique as a platform for high-throughput discovery.



Figure 3.1 SOLiD™ System.

1. *Probe specificity* is provided by degenerate bases at five positions (three are fully degenerate and two are partially degenerate, as they are linked to dye) within the dye-labeled oligonucleotide probes. Probes with sequences complementary to the unknown target are ligated to a universal primer in a stepwise fashion.
2. *Highly efficient chemical cleavage of the ligated probes* generates a 5' phosphate capable of participating in the next round of ligation.
3. *Dye-labeled sequencing products* are removed after five or seven rounds of ligation. This “reset” is followed by additional ligation cycles to a new universal primer positioned one nucleotide 5' from the previous universal primer. This process is repeated several times to complete the collection of sequence data.

3.2.1

The SOLiD Platform

3.2.1.1 Library Generation

The SOLiD System uses two types of randomly generated DNA sequencing libraries: a “fragment” library and a “mate-paired” library (Figure 3.2a). In the fragment library, the DNA to be sequenced is sheared (e.g., by sonication) to 60–90 bp fragments, the optimal size for amplification with this system. The adapters that are used for amplification are ligated to the ends of these fragments, and the ligation products are selected for emulsion PCR amplification.

The construction of a mate-paired library is more complex than that of a fragment library, it begins similarly with DNA sample shearing. The versatility of this protocol allows greater fragment size variability, 0.6–10 kb, depending on the application of interest. Fragments of the appropriate size are gel-purified and ligated to a synthetic primer, called a CAP adapter (Figure 3.2a). The capped fragments are diluted and ligated under conditions that favor recircularization over end-to-end ligation, resulting in a random library of circularized fragments of a predefined size (Figure 3.2a).

Paired ends are generated from these circularized molecules by digestion with the restriction enzyme EcoP15I. The recognition site for this enzyme is present at both ends of the CAP adapter. Such class III restriction enzymes cleave the DNA at a defined distance from their recognition sequence [1]. EcoP15I cuts 25 and 27 bases from its recognition site (Figure 3.2a). This digestion releases the intervening genomic DNA sequences from the 25–27 base pairs of the sequence that defined the ends of the original fragment. The random “mate-paired” fragments can then be purified from the intervening DNA by affinity capture, using a biotin tag designed into the CAP primer. This process results in a random set of paired 27-base DNA sequences that were originally separated in the genome by the selected fragment size (2–3 kb in this example) used to construct the library (Figure 3.2a). The same adapters that are used to flank the fragment library are now ligated to the mate-paired library, and primers complementary to adapter sequences are used for amplification of the library by emulsion PCR. Sequencing of mate-paired libraries has several advantages including downstream bioinformatics analysis [2, 3] for proper assembly of complex organisms and resolution of structural variations and successful detection of insertions, deletions, duplications, and rearrangements.

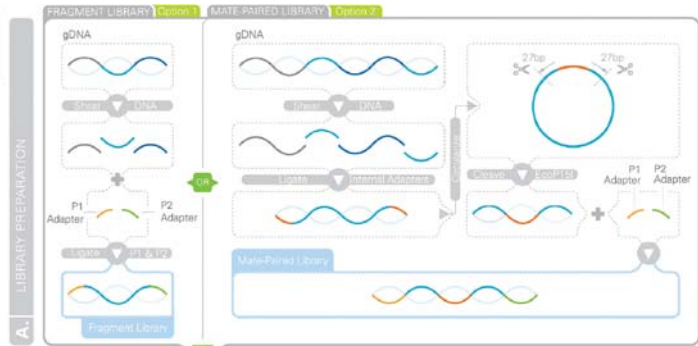
3.2.1.2 Emulsion PCR

The same emulsion PCR process is used for both library types, and it follows the protocol described by Dressman *et al.* [4] with minor modifications. A reverse-phase emulsion is prepared from specific oils and aqueous-based reagents that contains the library, PCR components, and 1 μm paramagnetic beads that have one of the PCR primers attached to their surface. For the purpose of clarity, the primer is hybridized to a ligated adapter, called P1 adapter, which is close to the magnetic bead (Figure 3.2b). PCR occurs in the aqueous microreactors formed when the emulsion is created (Figure 3.2b). Before the emulsion is made, the library in the aqueous phase is diluted to maximize the number of microreactors that will contain one DNA template and one magnetic bead. According to a normal Poisson distribution, a proportion of the microreactors will contain one bead and one template from the library, suitable for emulsion PCR (Figure 3.2b). During PCR amplification, the “productive” microreactors will drive the clonal amplification of individual templates from the library onto individual magnetic beads.

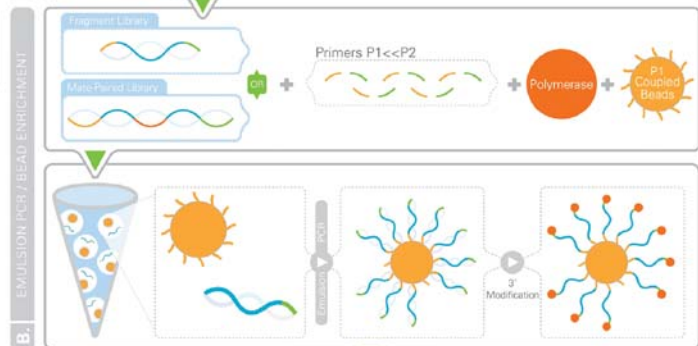
3.2.1.3 Bead Purification

The emulsions are broken to release the magnetic beads from the microreactors, using standard magnetic bead-purifying racks and successive washes with appropriate buffers. The magnetic beads are then mixed with large polystyrene capture beads that have a sequence complementary to the second PCR adapter (P2). Beads with extended PCR product will contain P2 sequences at their ends and will hybridize to the P2 complementary sequence on the surface of the polystyrene capture beads. The resulting bead complexes are purified using a glycerol gradient, and the magnetic beads with attached PCR products are eluted. This step routinely results in a total P2 + bead enrichment of more than 90% beads templated.

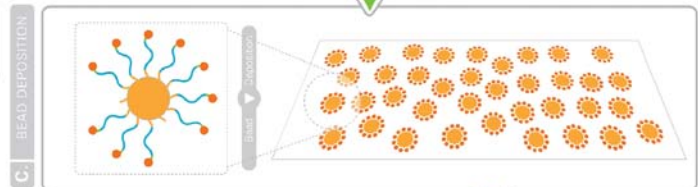
Panel A



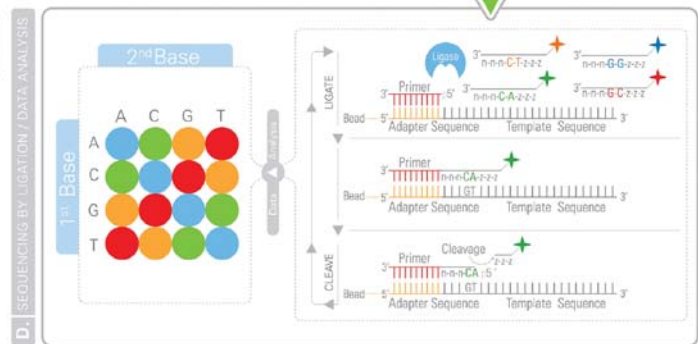
Panel B



Panel C



Panel D



3.2.1.4 Bead Deposition

The purified beads are then covalently attached to chemically modified glass slides through the chemically modified ends of the extended PCR-amplified templates (Figure 3.2c). Unlike picotitre plate production, in which a defined number of wells have been machined onto a plate, the only limitation to the number of beads that can be deposited on the surface of the slide is the diameter of the beads (1 μm). Currently, the SOLiD System can support 50,000 to 100,000 P2 + beads/ 0.75 mm^2 on an approximately 1350 mm^2 slide. However, the 1 μm diameter of the beads used in this system allows many more beads to be deposited on the slide surface. As the number of P2 + beads deposited on a slide increases, the number of bases that are generated from a slide increases accordingly with no reagents being used.

3.2.1.5 Sequencing by Ligation

The most novel aspect of the SOLiD System is the use of DNA ligase and fluorescently labeled oligonucleotide probes to perform the sequencing reaction. The basics of the sequencing chemistry are illustrated in Figure 3.2. Briefly, a universal sequencing primer complementary to the P1 adapter sequence anchors subsequent ligation reactions. A pool of fluorescently labeled octamer (eight-base) probes, containing all possible combinations of A, C, G, and T at positions 1–5 (designated as Ns), interrogates the sequence of the unknown template on each bead. Four of the possible 1024 probes in the pool available for hybridization and ligation are shown in Figure 3.2d. Only the probe homologous to the first five bases of the unknown template will be in the proper position to be ligated to the universal sequencing primer. Probes that hybridize to other regions of the DNA sequence are not substrates for ligase, because the enzyme can establish a phosphodiester linkage only between the adjacent 5' phosphate of one oligonucleotide and the 3' hydroxyl of the second oligonucleotide. The 3' end of the interrogating probe will then efficiently ligate to the 5' end of the universal sequencing primer (Figure 3.2d).

Figure 3.2 Schematic representation of the steps involved in SOLiD sequencing. (a) The SOLiD System uses both fragment and mate-paired libraries. The type of library used depends on the application and information needed. (b) Clonal bead populations are prepared in microreactors that contain template, PCR reaction components, beads, and primers. After PCR, the templates are denatured and bead enrichment is performed to separate beads with extended templates from undesired beads. The template on the selected beads undergoes a 3' modification to allow covalent bonding to the slide. (c) 3'-Modified beads are deposited onto a glass slide. A deposition chamber enables a slide to be segmented into 1, 4, or 8 chambers during

the bead-loading process. (d) Primers then hybridize to the adapter sequence, and four color dye-labeled probes compete for ligation to the sequencing primer. Specificity of probe ligation is achieved by interrogating every fourth and fifth base during ligation. Five to seven rounds of ligation, detection, and cleavage record the color at every fifth position. The number of rounds is determined by the type of library. Following each round of ligation, a new complementary primer, offset by one base in the 5' direction, is laid down for another ligation series. Five rounds of primer rest and ligation (five to seven ligation cycles per round) generate 25–35 bp of sequence for each tag. This process is repeated for the second tag during mate-paired sequencing.

Numerous publications have demonstrated the selectivity and specificity of ligation [5–7]. The probes are labeled with one of four fluorescent dyes, each associated within the probe with a distinct set of four dinucleotide combinations at positions 1 and 2 (Figure 3.2d). After illumination, the dye fluorescence is recorded with a xenon lamp. The dye attached to the ligated probe and the nucleotides at positions 6–8 must be removed before the next round of ligation. The dye and these nucleotides are removed by chemical cleavage of a modified linkage between probe nucleotides 5 and 6 (Figure 3.2d).

Depending on the desired read length, four to six additional rounds of ligation are performed in exactly the same manner as the first round. At the end of the first five (or seven) cycles of ligation, the colors of probes corresponding to possible dinucleotide combinations may be determined at positions 4 + 5, 9 + 10, 14 + 15, 19 + 20, 24 + 25, 29 + 30, and 34 + 35 from the end of the unknown template (Figure 3.2d). To collect the information about the remaining colors, the extended products resulting from the first five to seven ligations are chemically stripped from the template. The remaining colors are then determined by ligating the pools of labeled probes to additional unique universal sequencing primers that are offset from the original primer by 1, 2, 3, or 4 nucleotides (Figure 3.2d). This feature is known as “resetting” because it resets the ligation cycle (i.e., after seven ligations to the first primer, a new primer is utilized), the signal-to-noise ratio of the next ligation cycle to that of the first. After 7 ligation cycles are carried out for five rounds of ligation, sequencing of 35 bp of the unknown template will have occurred. Currently, sequencing read lengths of 25–35 bases are supported for fragment libraries, although read lengths of up to 60 bases have been demonstrated.

With mate-paired libraries, the first 25 cycles of ligation are performed in exactly the same manner as outlined for the fragment library sequencing reactions. Since the EcoP15I restriction enzyme cleaves 25–27 bases from its restriction site, the read length of the mate-paired fragments is 25 bases. Another 25 cycles of ligation are then performed in exactly the same manner by using a universal sequencing primer complementary to the internal CAP adapter (Figure 3.2a). This results in 25 bases of sequence from two mate-paired ends of templates that were originally separated by a known distance at the beginning of library preparation (2×25 bp).

The SOLiD System can generate in excess of 6 billion bases of DNA sequence per run if the entire surface (called a segment) of both slides is devoted to a single mate pair sample. An additional feature, sectioning of the slide into 4 or 8 segments, allows multiple samples to be run on the same slide by disposing beads from each sample onto a unique segment. At this time, it is possible to deposit up to eight samples on each slide or sixteen samples for the two slides. Additional multiplexing of samples on a slide has been demonstrated by incorporating a “barcode” into one of the adapters and performing additional ligation reactions to determine the sequence of the barcode. While multiplexing on the SOLiD System has not yet been commercialized, this advancement has the potential to permit multiple samples to be mixed in a single segment of the slide. The number of beads deposited per barcoded sample can be readily calculated if the number of bases needed for the application and the number of bases generated per bead are known.

3.2.1.6 Color Space and Base Calling

The SOLiD System uses a novel system based on two-base encoding that results in production of a sequence in “color space.” Each of the four possible dyes represents four potential dinucleotides combinations (Figure 3.2d). During the ligation chemistry, each base is interrogated twice. The color of the dye for each templated bead at each ligation cycle is recorded and stored digitally. The concept of representing a nucleotide sequence as a color sequence by using two-base encoding is “color space.”

The benefit of two-base encoding is that the unique design of the matrix (dinucleotide combinations and dye color; Figure 3.2d) enables measurement errors to be easily distinguished from true nucleotide polymorphisms. A measurement error occurs when a color call is incorrect and thus a single color space call will not match the color space translation of the reference sequence. A true polymorphism requires two adjacent color calls to change at the same time. This allows easy discrimination between a measurement error and a polymorphism. This feature confers a clear advantage over the single-base encoding found in DNA polymerase-based systems, which are unable to distinguish a measurement error from a polymorphism and thus requires higher coverage. The unrivaled overall accuracy rate of sequences generated by the SOLiD System exceeds 99.94%.

3.3

SOLiD™ System Applications

3.3.1

Large-Scale Resequencing

The SOLiD System is ideally suited to generate large amounts of data for genome resequencing projects. Depending on the size and complexity of the genome, as well as the application of interest, the investigator can construct either a fragment library or a mate-paired library. A fragment library requires fewer manipulations and no enzymatic cleavage (i.e., EcoP15I); therefore, it can contain sequences that are not represented in a mate-paired library. A mate-paired library, however, is superior for data assembly, especially for organizationally complex genomes. In some cases, construction and sequencing of both library types may be desirable to maximize sequence coverage.

The obvious advantage of massively parallel sequencing is its extremely uniform and deep coverage of most organisms with only one or a limited number of runs. A 2 megabases organism has been sequenced to a depth >200-fold with a single run on the SOLiD System using a precommercial protocol (unpublished data).

3.3.2

De novo Sequencing

Although it is not yet clear if short sequences can be used successfully for *de novo* sequencing of most organisms, this question will be addressed quickly when data

from the SOLiD System have been tested and the bioinformatics tools for assembly have been refined. However, by combining SOLiD sequencing with traditional Sanger-based sequencing, the properties of Sanger sequencing (very long, highly accurate reads) can be leveraged to carry out a cost-effective *de novo* sequencing. These sequences can be used to assemble a genome “scaffold” or backbone. Goldberg *et al.* [8] showed that sequences generated by Sanger-based instruments successfully closed a number of assembly gaps that remained after 5× coverage of a bacterial genome. Short reads from the SOLiD System especially when using mate-paired libraries can be used to close gaps in the scaffold. The massively parallel sequencing capability of the SOLiD System, combined with Sanger sequencing, facilitates the assembly of *de novo* sequences by producing short fragments (from shearing) that are more effective with difficult genomic regions. The sequencing of short reads eliminates bias in the regions of a genome that are difficult to sequence because they cannot be cloned in bacteria, a problem with traditional Sanger sequencing. These so-called unclonable regions presumably contain genes or sequences that are either directly toxic to the host cell or somehow interfere with normal cellular function. In addition, sequences such as GC-rich regions, homopolymers, and other simple repeats, difficult to sequence with polymerase-based sequencing chemistries, are more likely to be represented in sequence assemblies derived from the SOLiD System’s ligation-based sequencing chemistry, which does not add bases sequentially. This feature effectively eliminates the possibility of out-of-phase extensions that occur when homopolymers are sequenced by pyrosequencing chemistry [8].

3.3.3

Tag-Based Gene Expression

In addition to traditional sequencing applications, such as *de novo* sequencing and resequencing, the massively parallel sequencing capability of the SOLiD System performs genome-wide, tag-based sequencing applications. Several sequence tag-based gene expression applications have been developed including SAGE [9], Super-SAGE [10], CAGE [11], and 5′-SAGE [12]. These techniques involve short sequences that are unique to a specific RNA species. Methodologies exist to isolate these “tags,” manipulate them, and determine the tag sequence. The number of tags sequenced has been shown to be proportional to the number of mRNA molecules in the population. Because these tag-based methods do not require *a priori* knowledge of the genome sequence under study, they therefore serve as an alternative method for array-based gene expression, which requires DNA sequence information. Tag-based sequencing is extremely valuable for plants and other uncharacterized genomes that require transcription profiling.

Furthermore, highly sensitive tag-based sequencing is superior to array technology for measuring differences in gene expression levels because a greater number of tags can be sequenced (Figure 3.3). While existing sequence tag methods have been validated on traditional sequencing platforms and TaqMan-based methods, the data must also be validated on the SOLiD System and other next-generation instruments, before these systems can be adopted by the scientific community.

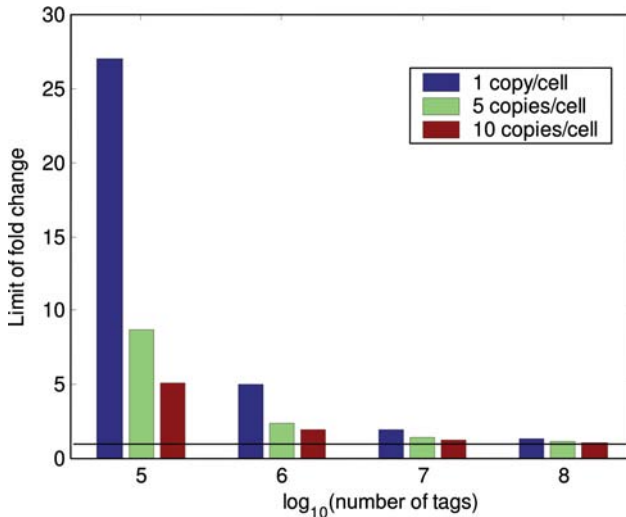


Figure 3.3 High sensitivity with tag-based sequencing. Theoretical limits of sensitivity detect transcripts of different copy numbers by means of sequence tags. The model shows that 107 sequence tags can detect even twofold changes in expression levels of single copy transcripts.

The number of sequence tags required for a 5'-SAGE gene expression experiment is normally in the range of 1×10^6 – 5×10^7 , depending on the application (I. Hashimoto, personal communication with Robert Nutter). As the SOLiD System randomly deposits beads containing clonally amplified templates, hundreds of thousands of $1 \mu\text{m}$ beads can be deposited on each square millimeter of the slide surface. The slide surface can also be physically separated into a number of different segments when the beads are deposited. If 2×10^6 sequence tags are needed for a sample, and 12 000 mappable beads or tags can be deposited onto each 0.75 mm^2 section of a slide, the area of a slide required to achieve the desired number of tags can be determined by the following calculation: $2\,000\,000/12\,000 \times 0.75$.

Because the number of tags required for an experiment comprises only a portion of a slide, various samples or biological controls can be run on a single slide. Configuring the slide appropriately lets you run different samples under the same sequencing conditions, providing a controlled experimental environment for your samples. As bead (tag) density increases per slide, a greater number of beads can be used in a single segment, which improves the sensitivity of gene expression analysis.

3.3.4

Whole Transcriptome Analysis

Whole transcriptome analysis of complex genomes is also enabled by the massively parallel sequencing capacity of the SOLiD System. Total, nonpolysomal RNA can be isolated, fragmented, and converted to cDNA flanked by P1 and P2 adapters. The

resulting sequence is then compared with the appropriate reference sequence to unambiguously map to transcribed regions. Sequence tags generated on the SOLiD System from mouse embryonic stem cell RNA uniquely identify >20 000 transcripts. The same cDNA, analyzed with Illumina's BeadArray technology, identifies ~9000 unique transcripts. Further analysis of the novel transcripts identified from SOLiD System sequence tags demonstrates that the majority are low-abundance transcripts present below the level of detection by standard arrays. The data further demonstrates the sensitivity of the SOLiD System. A further benefit from the SOLiD System sequencing data permits identification and quantification of splicing events [13]. The sequence data generated from each experiment must be stored and processed by software specifically designed for each application. Sequence tags, the mRNAs they correspond to, and their frequency in a sample must be tracked and tabulated. Serviceable analysis packages for each type of application are being developed by members of the scientific community, and Applied Biosystems is working to assure that these tools are freely available to researchers interested in developing the application in their laboratories. The amount of data generated by the SOLiD System's massively parallel sequencing technology presents significant bioinformatics challenges. The system generates hundreds of times more data than traditional sequencing (i.e., Sanger) systems. Primary analysis and sequence alignment tools are provided with the SOLiD System; however, downstream analysis software to manage the sequence data generated with the SOLiD System will need to be developed and distributed in a similar manner.

3.3.5

Whole Genome Resequencing

Large-scale resequencing projects need a highly parallel system to provide the depth of coverage required for variation detection. A highly accurate system is also critical to reduce false positive rates and provide the sensitivity necessary to detect mutations in pooled or heterogeneous samples. The SOLiD System enables a new level of whole genome sequencing and increased sample throughput, while requiring substantially less time and fewer resources than required by competing technologies.

3.3.6

Whole Genome Methylation Analysis

DNA methylation involves the regulation of many cellular processes, including X chromosome inactivation, chromosome stability, chromatin structure, embryonic development, and transcription. Methylation of specific genomic regions inactivates gene expression, while demethylation of other regions leads to inappropriate gene expression or chromosomal instability.

Bisulfite sequencing using automated capillary electrophoresis (CE) instruments is the gold standard for targeted analysis of the methylation status of specific regions; however, it does not scale to genome-wide analyses. Methylation patterns have been

studied by using a combination of enzymes that have differential sensitivity to CpG methylation. Samples are enriched for methylated regions and compared with samples depleted of methylated sequence [14]. Mate pair libraries have been made from normal tissue and a breast cancer cell line and sequenced on the SOLiD platform. Preliminary data generated from these libraries has shown differences in the methylation patterns. The SOLiD System provides an ultrahigh-throughput method for analyzing such samples with a much higher level of resolution than traditional methods.

3.3.7

Chromatin Immunoprecipitation

ChIP is a useful technique for identifying transcription factor-binding elements and characterizing their involvement in gene regulation. Historically, ChIP reactions were analyzed using IP Western blotting methods, and more recently by microarrays, also known as ChIP-on-Chip experiments [15]. Microarray technologies provide a method for global ChIP analysis, but their probe design is hypothesis-driven, has limited sensitivity, and is typically limited to known promoter regions. The SOLiD System's massively parallel sequencing has overcome this limitation and supports hypothesis-neutral ChIP sequencing, or ChIP-seq. The SOLiD System's ability to generate millions of sequence tags (read length: 35 bp) in a single run enables whole genome ChIP analysis of complex organisms. Sequence tags are then counted and mapped to a reference sequence to identify specific regions of protein binding. The system's ultrahigh throughput provides researchers with the sensitivity and statistical resolving power required to accurately characterize the protein/DNA interactions of an entire genome. Additionally, the system's flexible slide format permits the analysis of both normal and diseased samples in a single run.

3.3.8

MicroRNA Discovery

MircoRNAs (miRNAs) are a type of highly conserved small, non-coding RNA molecules (ncRNAs) encoded by many organisms, which play an important role in gene expression. Recent publications have shown that a large portion of the human genome is transcribed into miRNAs or other ncRNAs. Arrays have been the traditional method for studying small RNA expression. However, arrays are not good discovery platforms due to their limited dynamic range and the inability to have oligonucleotides on an array for RNAs you do not know are present. Sequencing of cDNA molecules made from small RNAs is ideal for discovery purposes. Once the RNA is converted to DNA, the sequence can readily be determined. Short (35–50 bp) cDNA sequences are sufficiently long to unambiguously map the location of the RNAs back to the genome. The dynamic range of the application is driven by the number of tags generated (Figure 3.3) and the SOLiD System is capable of generating hundreds of millions of short sequences tags

in each run. Recently, Applied Biosystems developed the SOLiD Small RNA Expression Kit which converts, in a single day, a few nanograms of RNA into cDNA libraries. The sequence of these tags is used to study expression levels of individual miRNAs and discover novel ncRNAs.

Experiments using this kit have already shown that the organization of miRNAs is very complex. Novel miRNAs have been predicted based on known miRNA structure and these miRNAs are currently being validated. Additionally, the presence of miRNA isoforms (isomiRs, RNA molecules having 5' and 3' ends that differ from the reference sequence) have been shown to be present, thereby complicating our understanding of the role that these molecules play in controlling gene expression. Clearly, the SOLiD System opens the doors to a more complete understanding of the role ncRNAs play in the control of gene expression in complex organisms.

3.3.9

Other Tag-Based Applications

As the cost of generating sequence data has decreased, it is becoming possible to move essentially all genetic analysis applications to sequence-based techniques. Even before the availability of next-generation sequencing, scientists described genome-wide genetic analysis applications that can be readily converted to sequence-based methods. While gene expression was the first to make use of sequence-specific tags, other applications, such as digital karyotyping (or genome-wide structural variation analysis) [16], end-sequence profiling (ESP) [17], and biomarker detection of microbial organisms [18], also use them. Although the technologies differ, they have been developed with one common goal that is to enrich regions of interest from the rest of the genome using various approaches. These applications can be easily converted to the SOLiD System via system-specific primers, clonal amplification, and sequencing. Application-specific data analysis tools must be developed that will permit enormous amount of raw sequence data to be organized efficiently. Developing data analysis tools capable of handling the massive amounts of sequence data that the SOLiD System generates is crucial.

3.4

Conclusions

The SOLiD System's massively parallel sequencing capability is revolutionizing genetic analysis by reducing costs and vastly increasing the amount of data that can be generated per sequencing experiment. Applications that use large numbers of sequence-specific tags, including gene expression, digital karyotyping, chromosome immunoprecipitation, and miRNA discovery, are now possible on a genome-wide scale. In addition, the SOLiD System enables researchers to develop and execute genome-wide genetic applications, increasing the amount of data generated and accelerating the discovery process to improve scientific endeavors.

Acknowledgments

Anna Berdine, Michael Gallad, Sue Ann Molero, Julie Moore, Michael Rhodes, Elizabeth Sanchez, Anjali Shah, and Carla Wike have reviewed and edited this chapter.

References

- 1 New England Biolabs, Restriction endonucleases overview; http://www.neb.com/nebecomm/tech_reference/restriction_enzymes/overview.asp (Accessed July 3, 2008).
- 2 Raphael, B., Volik, S., Collins, C. and Pevzner, P. (2003) Reconstructing tumor genome architectures. *Bioinformatics*, **19**, 162–171.
- 3 Whiteford, N., Haslam, N., Weber, G., Prügel-Bennett, A., Essex, J.W., Roach, P.L., Bradley, M. and Neylon, C. (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*, **33** (19), e171.
- 4 Dressman, D., Yan, H., Traverso, G., Kinzler, K. and Vogelstein, B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (15), 8817–8822.
- 5 Luo, J. and Barany, F. (1996) Identification of essential residues in *Thermus thermophilus* DNA ligase. *Nucleic Acids Research*, **24** (15), 3079–3085.
- 6 Liu, P., Burdzy, A. and Sowers, L. (2004) DNA ligases ensure fidelity by interrogating minor groove contacts. *Nucleic Acids Research*, **32** (15), 4503–4511.
- 7 Bhagwat, A., Sanderson, R. and Lindahl, T. (1999) Delayed DNA joining at 3' mismatches by human DNA ligases. *Nucleic Acids Research*, **27** (20), 4028–4033.
- 8 Goldberg, S.M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S.A., Lauro, F.M. *et al.* (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 11240–11245.
- 9 Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270** (5235), 368–369, 371.
- 10 Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Kruger, D.H. and Terauchi, R. (2003) Gene expression analysis of plant host–pathogen interactions by SuperSAGE. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15718–15723.
- 11 Shiraki, T., Kondo, S. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (26), 15776–15781.
- 12 Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S. and Matsushima, K. (2004) 5'-end SAGE for the analysis of transcriptional start sites. *Nature Biotechnology*, **22**, 1146–1149.
- 13 Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, **5** (7), 613–619.
- 14 Rollins, R.A., Haghghi, F., Edwards, J.R., Das, R., Zhang, M.Q., Ju, J. and

- Bestor, T.H. (2006) Large-scale structure of genomic methylation patterns. *Genome Research*, **16** (2), 157–163.
- 15 Cowell, J.K. and Hawthorn, L. (2007) The application of microarray technology to the analysis of the cancer genome. *Current Molecular Medicine*, **7** (1), 103–120.
- 16 Wang, T.L., Maierhofer, C., Speicher, M.R., Lengauer, C., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002) Digital karyotyping. *PNAS*, **99**, 16156–16161.
- 17 Volk, S., Zheo, S., Chin, K., Brekner, J.H., Herndon, D.R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W.L., Magrane, G., de Jong, P., Gray, J.W. and Collins, C. (2002) End-sequence profiling: sequence-based analysis of aberrant genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 7696–7701.
- 18 Tengs, T., LaFramboise, T., Den, R., Hayes, D., Zhang, J., DebRoy, S., Gentleman, R., O'Neill, K., Birren, B. and Meyerson, M. (2004) Genomic representations using concatenates of type IIB restriction endonuclease digestion fragments. *Nucleic Acids Research*, **32**, e121–e129.

4

The Next-Generation Genome Sequencing: 454/Roche GS FLX*Lei Du and Michael Egholm*

4.1

Introduction

DNA sequencing technology has been undergoing rapid change in recent years. Since Frederick Sanger and coworkers developed the enzyme-based chain-termination method in 1975, we have seen significant optimization on the technology along with the development and commercialization of automated DNA sequencers. Next-generation sequencing refers to a new class of instrumentation that combines rapid sample preparation and dramatically enhanced total throughput and promises to bring unprecedented capability in genomic research. With current state-of-the-art Sanger method, sequencing a large DNA molecule (e.g., whole bacterial genome) involves shearing of DNA into fragments, shotgun cloning in plasmid vectors, growth and purification in bacteria, and sequencing on a 96- or 384-lane automated capillary sequencer. The entire process takes a minimum of 4–6 weeks in a fully automated large-scale facility, with significant investment in robotic hardware, consumables, and human labor.

In 2004, 454 Life Sciences (now part of Roche Diagnostics) commercialized the first next-generation sequencing instrument, the Genome Sequencer 20 [1]. It utilizes emulsion-based cloning and PCR amplification with miniaturized, pyrophosphate-based sequencing using beads on a PicoTiterPlate™ (PTP). The PTP is made of fused optical fibers with chemically etched wells on one end of the fiber. Light generated during the course of pyrosequencing is conducted by way of total internal reflection from the well through the bottom of the fiber and detection by a juxtaposed astronomical-grade CCD camera. The parallel nature of this technology allows simultaneous measurement of more than 200 000 DNA sequences totaling 20–40 million base pairs. In December 2006, 454 Life Sciences introduced the second-generation instrument, Genome Sequencer FLX (GS FLX), with a throughput of 100 million base pairs and an average read length of more than 240 bases. Improvement in total throughput was achieved by a combination of optimized reaction chemistry, image processing, and instrument hardware. With this level of

throughput and read length, researchers can now conduct experiments that were not possible previously due to the prohibitive cost and long project duration of traditional sequencing technology. Through April 2008, more than 200 peer-reviewed articles have been published, demonstrating the utility of 454 Sequencing in a broad range of research applications. A select subset of these applications is listed in the “References” section, such as microbial genomics and drug resistance [2–9], plant genetics [10–12], PCR amplicons [42–44], small RNA [13–28] gene regulation [29–33], transcriptome analysis [34–41], metagenomics and environmental diversity [45–51], ancient DNA analysis [52–55], and human genome sequencing [56, 57].

4.2

Technology Overview

The GS FLX system has three main components: DNA library preparation, emulsion PCR, and PicoTiterPlate™ sequencing. The library preparation step (Figure 4.1) generates a pool of single-stranded DNA fragments, each carrying universal adapters, A and B, one on each end of the molecule. The universal adapters are 44 bases in length and are composed of a 20-base PCR primer, a 20-base sequencing primer, and a 4-base key sequence for read identification and signal normalization. The DNA fragments can be generated from mechanical shearing of a long DNA molecule such as whole genome or long PCR product and subsequently ligated with the universal

**The genome sequencer FLX system provides a complete solution-
from sample preparation to digital data analysis.**

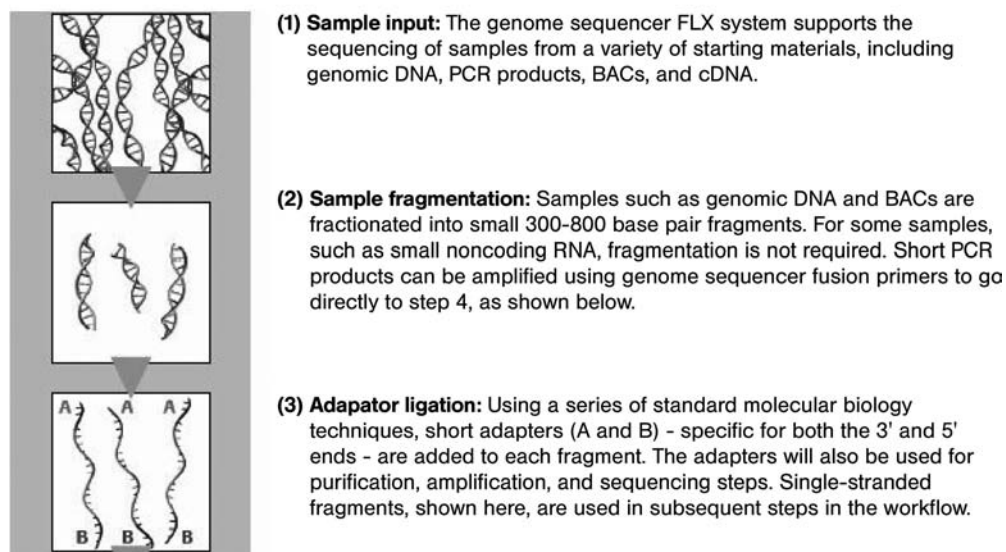


Figure 4.1 DNA sample preparation.

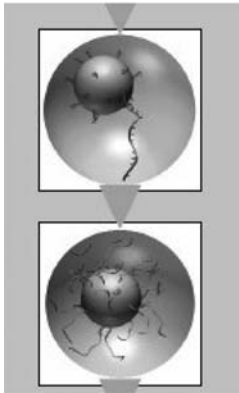


Figure 4.2 Emulsion PCR.

(1) One fragment = One bead: The first step in emulsion PCR (emPCR) is shown. The adapters enable hundreds of thousands of single-stranded fragments to bind to their own unique beads. The beads are then encapsulated into individual droplets formed by a water-in-oil emulsion, creating a microreactor containing one bead with one unique fragment. Each unique fragment is amplified without the introduction of competing or contaminating sequences. The entire fragment collection is amplified in parallel.

(2) One bead = One read: The emPCR process amplifies each fragment to a copy number of several million per bead. Subsequently, the emulsion is broken while the fragments remain bound to their specific beads. After enrichment, the clonally amplified bead is ready to load onto the PicoTiterPlate device for sequencing.

adapters at each end. Alternatively, input DNA can be generated by targeted PCR amplification of genes and loci of interest using site-specific primers with 5' adapter overhangs. No adapter ligation is necessary for this “amplicon library” method.

For emulsion PCR, the library material is mixed at limited dilution with beads (about 35 μm in size) carrying one of the adapter primers on its surface in a water-in-oil emulsion setup (Figure 4.2). The aqueous phase of the emulsion contains the bead, PCR primers (with the same sequence as the universal adapters and one of the primers having a biotin at the 5' end), nucleotides, and polymerase. Through limited dilution, each bead will encounter mostly zero and sometimes one DNA molecule prior to PCR. The chance of having multiple DNA fragments in one emulsion droplet enclosure is very small and reads from such droplets can be subsequently removed in the data processing step. The emulsion mix undergoes normal thermal cycling, leading to clonal amplification of the single-molecule DNA template in each enclosure and population of that bead with millions of identical molecules. At the end of the PCR process, emulsions are solubilized and template-carrying beads are enriched using magnetic streptavidin beads that select biotin-containing beads. The bead-bound double-stranded DNA fragments are rendered single stranded by denaturing, and sequencing primer is annealed to the free 3' end of the millions of clonal fragments on each bead. The beads are deposited into the wells of the PTP and overlaid with beads carrying sulfurylase and luciferase, enzymes needed for the pyrosequencing reaction.

For sequencing on the GS FLX, the PTP is inserted into a flow chamber on the instrument and placed directly in front of the CCD camera (Figure 4.3). Nucleotides are flown across the open surface of the PTP via the fluidics subsystem inside the sequencer, and DNA synthesis is carried out in real time in each well. Positive incorporation of one or more nucleotides at a given flow of a particular nucleotide will generate free pyrophosphate, which is converted to ATP by sulfurylase. ATP subsequently drives the oxidation of luciferin by luciferase and light is emitted in a stoichiometric fashion. The emitted photons are captured through the bottom of the PTP well by a CCD camera. The light intensity is proportional to the number of

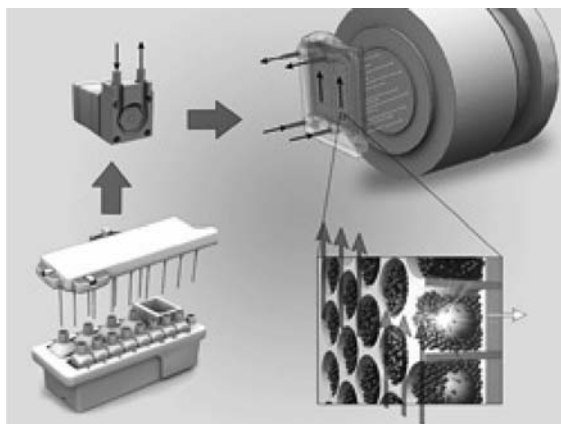


Figure 4.3 Sample loading into PicoTiterPlate and FLX instrument.

bases incorporated and the signal transduction enzymatic activity in each well (e.g., a stretch of four Gs will generate four times the light than that from a single G nucleotide when C is flown). As the amount of enzyme activity in each well is dictated by the number of enzyme beads deposited in each well, and the four-base key sequence at the beginning of each polymerase reaction is a fixed stretch of monomer nucleotides, the overall signal is normalized by the light intensity generated during the sequencing of the key for each well (Figure 4.4). The data processing software pipeline includes image processing, signal processing and normalization, noise reduction and phase correction, and base calling. The GS FLX system also comes with graphical user interface software to allow browsing of instrument run data, quality monitoring, as well as tools for *de novo* assembly and remapping against reference to identify mutations.

The original GS20 system was published with a detailed protocol described in the “Supplemental Materials” section [1]. The GS FLX system has the same core

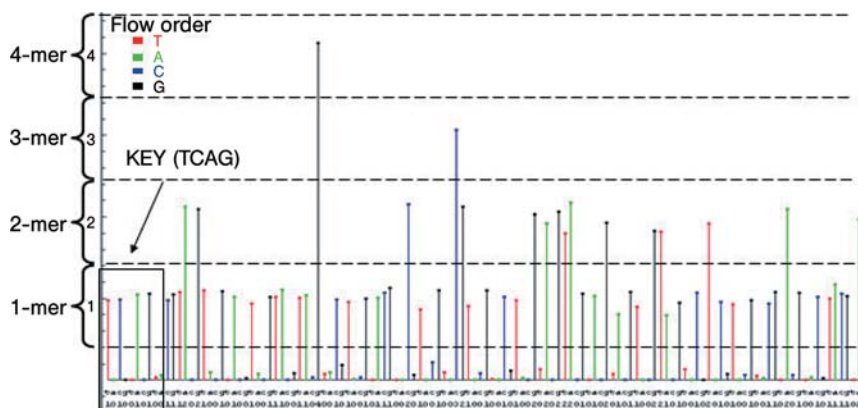


Figure 4.4 Signal flowgram and base calling.

components as GS20, with a series of improvements in sample preparation, sequencing, instrument run, and bioinformatics. The detailed protocol for DNA sequencing on the GS FLX is described in the following section, and references can also be found online (<https://www.roche-applied-science.com/sis/sequencing/flx/index.jsp>).

4.3

Software and Bioinformatics

Depending on the research application, the samples used, and the amount of sequence data generated, there are a variety of data analysis methods that will help manage and interpret results from the GS FLX system. Three main methods are described here, which include whole genome assembly, whole genome mapping and mutation detection, and PCR-based ultradeep sequencing.

4.3.1

Whole Genome Assembly

A bacterial whole genome assembly typically requires sequencing the organism to a 20× depth (ratio of total bases generated over the length of the genome), followed by running 454 Newbler software. Newbler constructs *de novo* assemblies of the reads from one or more sequencing runs and generates a set of contigs and consensus sequence for each contig. An option allows the inclusion of mate pair sequencing data into the analysis, to help orient the assembled contigs into scaffolds. Mate pair sequencing requires a separate sample preparation step and the protocol can be found online as described earlier. The accuracy of *de novo* assembly has been steadily improving and can reach beyond 99.999% at consensus level (Figure 4.5a).

4.3.2

Resequencing and Mutation Detection

Data from whole genome shotgun sequencing can be used to map against a reference genome with highly homologous sequence, and individual mutations can be detected by comparing the consensus base call with the reference (both homozygous and heterozygous SNPs and indels). The tool for this analysis is the 454 Reference Mapper (Figure 4.5b).

4.3.3

Ultradeep Sequencing

Using the amplicon library procedure, targeted genes and chromosomal regions can be selectively amplified and sequenced to a very high depth. This will allow the detection of variants with frequency as low as 0.5% in a pool of heterogeneous variants. Applications include sequencing tumor DNA to detect rare somatic



Figure 4.5 (a) Screenshot from 454 *de novo* assembler. (b) Screenshot from 454 Reference Mapper, showing SNP position of A/G.

mutations, sequencing disease-associated regions from individuals or pools for gene discovery and genotyping, and viral sequencing for drug resistance characterization. The Amplicon Variant Analyzer software is a comprehensive and user-friendly package that was developed to facilitate such data analysis. Figure 4.6a shows an example of a 15-base deletion occurring at 3% frequency from about 300 aligned reads, and in Figure 4.6b, a single-base substitution of A to G is detected at 6% frequency with 60 reads.



Research Applications

Andries and colleagues published a study [2] describing the discovery and validation of a point mutation in the ATP synthase gene in *Mycobacterium tuberculosis* conferring resistance to a potent small molecule inhibitor R207910. This experiment

was a classic example of how whole genome shotgun sequencing and mutation detection can be used to sequence and compare multiple strains of a bacterial organism, where one strain is sensitive to drug treatment and two strains are resistant (either by natural occurrence or by laboratory selection). Four single-base mutations were identified that were shared by the resistant strains and were absent in the sensitive strain. The GS system has since been rapidly adopted as a robust and cost-effective method to analyze whole bacterial genomes [2–9].

Plant geneticists took advantage of their existing tiled BAC libraries and sequenced them in pools [10–12], thus reducing the complexity of assembly from gigabase-sized genome to a few megabases.

One research area that has witnessed rapid advancement in the last few years is the discovery and characterization of small RNA molecules in model organisms [13–28]. Small RNAs are known to play important roles in regulating RNA stability, protein synthesis, chromatin structure, and genome organization. Henderson *et al.* [13] studied microRNA patterns from four dicer gene mutants in *Arabidopsis thaliana* and characterized their enzymatic function in small RNA production. Girard and colleagues identified a new class of small RNA called piRNA from mouse, and Lau *et al.* did the same for rat [14, 15]. Hannon and colleagues have since published a series of articles on the study of piRNA in *Drosophila* and zebrafish [24, 27], as did Bartel and colleagues in the study of *Arabidopsis* and *C. elegans* [18, 21].

In the area of environmental sequencing and molecular diversity studies, the GS system has been shown to be very effective [45–51]. Sogin and colleagues sequenced the V6 hypervariable region of ribosomal RNAs and demonstrated that bacterial diversity estimates for the diffuse flow vents of Axial Seamount and the deep water masses of the North Atlantic are much greater than any published description of marine microbial diversity at the time [46]. Other metagenomic efforts include assessing microbial diversity in deep mines [47], soil [48], the virome in multiple ocean locations [49], and the relationship between microbial communities in animal gut and obesity [50].

The sequencing of large eukaryotic genomes such as human and mouse will continue to challenge the technology as the requirement for whole genome coverage exceeds tens of gigabases of raw data. As an intermediate step toward the ultimate whole human/animal genome sequencing, genes and disease association regions can be selectively amplified and sequenced in the GS FLX system using the amplicon protocol [42–44]. This protocol can facilitate two types of genetic interrogation on large genomes: (a) ultradeep sequencing and detection of rare alleles such as those found in tumors and (b) parallel sequencing of a large pool of amplicons covering many genes of interest. As a demonstration, Thomas and colleagues who published their work on the profiling of EGFR mutations in human nonsmall-cell lung carcinoma were able to detect low abundance mutations that are invisible to traditional Sanger technology [42].

A powerful method of studying large eukaryotic genomes is to analyze the total transcriptome [34–41]. This method reduces the sequencing complexity by several 100-folds and focuses the resources on the most important aspect of genome

analysis. Published methods include a modification of short tags [34, 37] or direct sequencing of full-length cDNA [36, 38, 40], with the ultimate goal of discovering novel transcripts, splice variants, and expression profiles.

A more recently developed approach for reducing genome complexity is to use microarrays as enrichment medium to select genes or genomic regions of interest and then subject the DNA collection to shotgun sequencing. Preliminary data show a great promise in such combination [56] with applications in population sequencing, disease association study, and candidate gene sequencing.

Recently, paired end sequencing was used to measure large-scale structural variations in the human genome [57]. Many previously undetected variations became visible due to high resolution and sufficient depth of coverage. The sequencing protocol contains a step where whole genomic DNA is sheared, end adapted, and ligated to form a self-circle. This circular DNA is then sheared and the two ends of DNA attached to the adapter linker are inserted into emulsion PCR and sequenced, generating two reads that span the genome at approximately 3 kb distance.

These mate-pair reads are very useful in linking *de novo* assembly contigs to form scaffolds. For example, an *E. coli* K12 genome with 20× oversampling can be assembled typically to about 100 contigs; with mate pairs, these will further be linked to produce 10 scaffolds. With further improvement in the assembly software, it will be possible to reconstruct repetitive segments in the genome and construct very long-range genomic contigs, ultimately achieving full assembly with only draft shotgun data.

The promise of routine human sequencing cannot be underestimated as the technology continues to undergo rapid improvement. Individuals will be able to obtain their own genetic blueprint and assess risks relative to disease, environment, and prescription drugs. A new generation of functional genomic studies can be carried out where an organism or group of organisms can be analyzed in parallel by studying their genome, transcriptome, chromosomal modification, and individual variations. Personalized medicine will become reality when sequencing and interpreting individual genomes become efficient and affordable.

References

Sequencing Technology

- 1 Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Whole Genome Sequencing

- 2 Andries, K., Verhasselt, P., Guillemont, J., Gohlmann, H.W., Neefs, J.M., Winkler, H., Van Gestel, J., Timmerman, P., Zhu, M., Lee, E., Williams, P., de Chaffoy, D., Huitric, E., Hoffner, S., Cambau, E., Truffot-Pernot, C., Lounis, N. and Jarlier, V. (2005) A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science*, **307**, 223–227.
- 3 Velicer, G.J., Raddatz, G., Keller, H., Deis, S., Lanz, C., Dinkelacker, I. and Schuster, S.C. (2006) Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 8107–8112.
- 4 Goldberg, S.M., Johnson, J., Busam, D., Feldblyum, T., Ferreira, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S.A., Lauro, F.M., Li, K., Rogers, Y.H., Strausberg, R., Sutton, G., Tallon, L., Thomas, T., Venter, E., Frazier, M. and Venter, J.C. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 11240–11245.
- 5 Hofreuter, D., Tsai, J., Watson, R.O., Novik, V., Altman, B., Benitez, M., Clark, C., Perbost, C., Jarvie, T., Du, L. and Galán, J.E. (2006) Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infection and Immunity*, **74**, 4694–4707.
- 6 Jung, D.O., Kling-Backhed, H., Giannakis, M., Xu, J., Fulton, R.S., Fulton, L.A., Cordum, H.S., Wang, C., Elliott, G., Edwards, J., Mardis, E.R., Engstrand, L.G. and Gordon, J.I. (2006) The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 9999–10004.
- 7 Smith, M.G., Gianoulis, T.A., Pukatzki, S., Mekalanos, J.J., Ornston, L.N., Gerstein, M. and Snyder, M. (2007) New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. *Genes & Development*, **21**, 601–614.
- 8 Highlander, S.K., Hulten, K.G., Qin, X., Jiang, H., Yerrapragada, S., Mason, E.O., Shang, Y., Williams, T.M., Fortunov, R.M., Liu, Y., Igboeli, O., Petrosino, J., Tirumalai, M., Uzman, A., Fox, G.E., Cardenas, A.M., Muzny, D.M., Hemphill, L., Ding, Y., Dugan, S., Blyth, P.R., Buhay, C.J., Dinh, H.H., Hawes, A.C., Holder, M., Kovar, C.L., Lee, S.L., Liu, W., Nazareth, L.V., Wang, Q., Zhou, J., Kaplan, S.L. and Weinstock, G.M. (2007) Subtle genetic changes enhance virulence of methicillin resistant and sensitive *Staphylococcus aureus*. *BMC Microbiology*, **7**, 99, 10.1186.
- 9 Hogg, J.S., Hu, F., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J.C. and Ehrlich, G.D. (2007) Characterization and modeling of the *Haemophilus influenzae* core- and supra-genomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biology*, **8**, R103.

BACs/Plastids/Mitochondria

- 10 Wicker, T., Schlagenhauf, E., Graner, A., Close, T.J., Keller, B. and Stein, N. (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*, **7**, 275–295.
- 11 Moore, M.J., Dhingra, A., Soltis, P.S., Shaw, R., Farmerie, W.G., Foltá, K.M. and Soltis, D.E. (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology*, **6**, 17–29.
- 12 Cai, Z., Penaflor, C., Kuehl, J.V., Leebens-Mack, J., Carlson, J.E., dePamphilis, C.W., Boore, J.L. and Jansen, R.K. (2006) Complete plastid genome sequences of *Drimys*, *Liriodendron*, and *Piper*:

implications for the phylogenetic relationships of magnoliids. *BMC Evolutionary Biology*, **6**, 77–96.

Small RNA

- 13 Henderson, I.R., Zhang, X., Lu, C., Johnson, L., Meyers, B.C., Green, P.J. and Jacobsen, S.E. (2006) Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nature Genetics*, **38**, 721–725.
- 14 Girard, A., Sachidanandam, R., Hannon, G.J. and Carmell, M.A. (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, **442**, 199–202.
- 15 Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P. and Kingston, R.E. (2006) Characterization of the piRNA complex from rat testes. *Science*, **313**, 363–367.
- 16 Lu, C., Kulkarni, K., Souret, F.F., MuthuValliappan, R., Tej, S.S., Poethig, R.S., Henderson, I.R., Jacobsen, S.E., Wang, W., Green, P.J. and Meyers, B.C. (2006) MicroRNAs and other small RNAs enriched in the *Arabidopsis* RNA-dependent RNA polymerase-2 mutant. *Genome Research*, **16**, 1276–1288.
- 17 Qi, Y., He, X., Wang, X.J., Kohany, O., Jurka, J. and Hannon, G.J. (2006) Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature*, **443**, 1008–1012.
- 18 Axtell, M.J., Jan, C., Rajagopalan, R. and Bartel, D.P. (2006) A two-hit trigger for siRNA biogenesis in plants. *Cell*, **127**, 565–577.
- 19 Berezikov, E., Thuemmler, F., van Laake, L.W., Kondova, I., Bontrop, R., Cuppen, E. and Plasterk, R.H.A. (2006) Diversity of microRNAs in human and chimpanzee brain. *Nature Genetics*, **38**, 1375–1377.
- 20 Pak, J. and Fire, A. (2007) Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science*, **315**, 241–244.
- 21 Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H. and Bartel, D.P. (2006) Large-scale sequencing reveals 21U-RNAs and additional MicroRNAs and endogenous siRNAs in *C. elegans*. *Cell*, **127**, 1193–1207.
- 22 Rajagopalan, R., Vaucheret, H., Trejo, J. and Bartel, D.P. (2006) A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes & Development*, **20**, 3407–3425.
- 23 Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangel, J.L. and Carrington, J.C. (2007) High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of MIRNA genes. *PLoS ONE*, **2**, e219–e232.
- 24 Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R. and Hannon, G.J. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, **128**, 1089–1103.
- 25 Zhang, X., Henderson, I.R., Lu, C., Green, P.J. and Jacobsen, S.E. (2007) Role of RNA polymerase IV in plant small RNA metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 4536–4541.
- 26 Howell, M.D., Fahlgren, N., Chapman, E.J., Cumbie, J.S., Sullivan, C.M., Givan, S.A., Kasschau, K.D. and Carrington, J.C. (2007) Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in *Arabidopsis* reveals dependency on miRNA- and tasiRNA-directed targeting. *The Plant Cell*, **19**, 926–942.
- 27 Houwing, S., Kamminga, L.M., Berezikov, E., Cronenbold, D., Girard, A., van den Elst, H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B., Plasterk, R.H.A., Hannon, G.J., Draper, B.W. and Ketting, R.F. (2007)

A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. *Cell*, **129**, 69–82.

- 28 Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K. and Hannon, G.J. (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science*, **316**, 744–747.

Chromosome Structure

- 29 Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., Green, R.D. and Dekker, J. (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research*, **16**, 1299–1309.
- 30 Johnson, S.M., Tan, F.J., McCullough, H.L., Riordan, D.P. and Fire, A.Z. (2006) Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Research*, **16**, 1505–1516.
- 31 Albert, I., Mavrich, T.N., Tomsho, L.P., Qi, J., Zanton, S.J., Schuster, S.C. and Pugh, B.F. (2007) Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature*, **446**, 572–576.
- 32 Nagel, S., Scherr, M., Kel, A., Hornischer, K., Crawford, G.E., Kaufmann, M., Meyer, C., Drexler, H.G. and MacLeod, R.A.F. (2007) Activation of TLX3 and NKX 2-5 in t(5;14)(q35;q32) T-cell acute lymphoblastic leukemia by remote 3'-BCL11B enhancers and coregulation PU.1 and HMGA1. *Cancer Research*, **67**, 1461–1471.
- 33 Bhinge, A.A., Kim, J., Euskirchen, G.M., Snyder, M. and Iyer, M.R. (2007) Mapping the chromosomal targets of STAT1 by sequence tag analysis of genomic enrichment (STAGE). *Genome Research*, **17**, 910–916.

Transcriptomes

- 34 Ng, P., Tan, J.J., Ooi, H.S., Lee, Y.L., Chiu, K.P., Fullwood, M.J., Srinivasan, K.G., Perbost, C., Du, L., Sung, W.K., Wei, C.L. and Ruan, Y. (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Research*, **34**, e84–e93.
- 35 Gowda, M., Li, H., Alessi, J., Chen, F., Pratt, R. and Wang, G.L. (2006) Robust analysis of 5'-transcript ends (5'-RATE): a novel technique for transcriptome analysis and genome annotation. *Nucleic Acids Research*, **34**, e126.
- 36 Bainbridge, M.N., Warren, R.L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., Mardis, E.R., Sadar, M.D., Siddiqui, A.S., Marra, M.A. and Jones, S.J. (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, **7**, 246–256.
- 37 Nielsen, K.L., Høgh, A.L. and Emmersen, J. (2006) DeepSAGE – digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Research*, **34**, e133.
- 38 Cheung, F., Haas, B.J., Goldberg, S.M., May, G.D., Xiao, Y. and Town, C.D. (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics*, **7**, 272–281.
- 39 Emrich, S.J., Barbazuk, W.B., Li, L. and Schnable, P.S. (2006) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, **17**, 69–73.
- 40 Weber, A.P., Weber, K.L., Carr, K., Wilkerson, C. and Ohlrogge, J.B. (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiology*, **144**, 32–42.
- 41 Toth, T.D., Varala, K., Newman, T.C., Miguez, F.E., Hutchison, S.K., Willoughby, D.A., Simons, J.F., Egholm, M., Hunt, J.H., Hudson, M.E. and Robinson, G.E. (2007)

Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science*, **318** (5849), 441–444.

Amplicons

- 42 Thomas, R.K., Nickerson, E., Simons, J.F., Jänne, P.A., Tengs, T., Yuza, Y., Garraway, L.A., LaFramboise, T., Lee, J.C., Shah, K., O'Neill, K., Sasaki, H., Lindeman, N., Wong, K.K., Borrás, A.M., Gutmann, E.J., Dragnev, K.H., DeBiasi, R., Chen, T.H., Glatt, K.A., Greulich, H., Desany, B., Lubeski, C.K., Brockman, W., Alvarez, P., Hutchison, S.K., Leamon, J.H., Ronan, M.T., Turenchalk, G.S., Egholm, M., Sellers, W.R., Rothberg, J.M. and Meyerson, M. (2006) Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature Medicine*, **12**, 852–855.
- 43 Binladen, J., Gilbert, M.T., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R. and Willerslev, E. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.
- 44 Dahl, F., Stenberg, J., Fredriksson, S., Welch, K., Zhang, M., Nilsson, M., Bicknell, D., Bodmer, W.F., Davis, R.W. and Ji, H. (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 9387–9392.
- Alexander, E.C. and Rohwer, F. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**, 57.
- 46 Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., Arrieta, J.M. and Herndl, G.J. (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115–12120.
- 47 Krause, L., Diaz, N.N., Bartels, D., Edwards, R.A., Puhler, A., Rohwer, F., Meyer, F. and Stoye, J. (2006) Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics*, **22**, e281–e289.
- 48 Leininger, S., Urlich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G.W., Prosser, J.I., Schuster, S.C. and Schleper, C. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, **442**, 806–809.
- 49 Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J.M., Mueller, J.E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C.A. and Rohwer, F. (2006) The marine viromes of four oceanic regions. *PLoS Biology*, **4**, e368.
- 50 Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R. and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.
- 51 Cox-Foster, D.L., Conlan, S., Holmes, E.C., Palacios, G., Evans, J.D., Moran, N.A., Quan, P.-L., Brieseman, T., Hornig, M., Geiser, D.M., Martinson, V., vanEngelsdorp, D., Kalkstein, A.L., Drysdale, A., Hui, J., Zhai, J., Cui, L., Hutchison, S.K., Simons, J.F., Egholm, M., Pettis, J.S. and Lipkin, W.I. (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*, **319** (5864), 724–725.

Metagenomics and Microbial Diversity

Ancient DNA

- 52 Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R.D., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., Ramm, M., Miller, W. and Schuster, S.C. (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, **311**, 392–394.
- 53 Gilbert, M.T., Binladen, J., Miller, W., Wiuf, C., Willerslev, E., Poinar, H., Carlson, J.E., Leebens-Mack, J.H. and Schuster, S.C. (2006) Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. *Nucleic Acids Research*, **35**, 1–10.
- 54 Stiller, M., Green, R.E., Ronan, M., Simons, J.F., Du, L., He, W., Egholm, M., Rothberg, J.M., Keates, S.G., Ovodov, N.D., Antipina, E.E., Baryshnikov, G.F., Kuzmin, Y.V., Vasilevski, A.A., Wuenschell, G.E., Termini, J., Hofreiter, M., Jaenicke-Despres, V. and Pääbo, S. (2006) Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 13578–13584.
- 55 Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M. and Pääbo, S. (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature*, **444**, 330–336.

Human Genome

- 56 Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., Weinstock, G.M. and Gibbs, R.A. (2007) Direct selection of human genomic loci by microarray hybridization. *Nature Methods*, **4** (11), 903–905.
- 57 Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Palejev, K., Carriero, N.J., Du, L., Taillon, B.E., Chen, C., Tanzer, A., Saunders, E., Chi, J., Yang, F.T., Carter, N.P., Hurles, M.E., Weissman, S., Harkins, T.H., Gerstein, M.B., Egholm, M. and Snyder, M. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

5

Polony Sequencing: History, Technology, and Applications

Jeremy S. Edwards

5.1

Introduction

The incredible success of the Human Genome Project (HGP) clearly illustrates how early investments in developing cost-effective methods of rapid DNA sequencing can have tremendous payoffs for the biomedical community. Over the course of a decade, through refinement, parallelization, and automation of established sequencing technologies, the HGP motivated a 100-fold reduction of sequencing costs, from \$10 per finished base to \$0.10 per finished base [1]. Now, in the wake of the HGP, the utility and potential impact of high-throughput DNA sequencing are even greater. I strongly believe that the completion of the human genome project marks the end of the beginning, rather than the beginning of the end, of the era of high-throughput sequencing. “Next-generation” sequencing technologies are providing faster, cheaper, and higher quality sequence, and as these technologies become more widely available they will likely have a tremendous impact on many areas of medicine and medical research.

Of the numerous next-generation sequencing approaches, polony sequencing is unique because it is an academically driven effort and, therefore, all the equipment is readily available and all the software is completely open source. Therefore, for many research groups and scientific applications, polony sequencing is likely the ideal method for many large-scale sequencing projects due to the accessibility of the technology. Polony sequencing is a versatile approach that can be modified for a number of applications (i.e., genome sequencing, mRNA tag library sequencing, etc.). In addition, polony sequencing is competitive with other approaches in terms of accuracy and is significantly cheaper. In this chapter, I will describe the history and current status of polony sequencing technology.

5.2

History of Polony Sequencing

High-throughput technologies for DNA sequencing have in general succeeded by spatially and temporally increasing the amount of data that can be obtained.

Next-Generation Genome Sequencing: Towards Personalized Medicine. Edited by Michal Janitz
Copyright © 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-32090-5

Typically, the spatial and temporal improvements have occurred through miniaturization of the processes and/or rapid sample processing. The development of polony technology is an extreme example of spatial compression; a polony array essentially consists of millions of distinguishable, immobilized, and femtoliter-scale “test tubes” filled with clonal DNA arising from individual DNA or RNA molecules via a single polymerase chain reaction (PCR). The fact that polony technology utilizes only a single microscope slide leads this technology to replace the complex robotics required to handle the tens-of-thousands of cloning and PCRs that feed conventional high-throughput sequencing. In this section, I will discuss how polony sequencing was practiced in the late 1990s and how this concept has led to an innovative sequencing technology that now has the potential for use in sequencing the human genome. The current status of polony sequencing technology is only a distant relative of the classical “polony” technology; however, I will briefly describe the original polony approach because it was the predecessor of the current polony sequencing. The original polony technology is also of historical importance because all the methods utilized for assaying and sequencing the DNA are identical to the methods used on the original polonies.

5.2.1

Introduction to Polonies

Polonies are conceptually related to bacterial colonies on an agar plate. The idea of plating bacterial cells on an agar gel at the appropriate dilution is to allow the formation of individual bacterial colonies. All individual cells in a colony are clones of the original cell that founded the colony, and all colonies are distinct. A polony (or polymerase colony) is a colony of DNA that is amplified from a single nucleic acid molecule within an acrylamide gel (Figure 5.1). Since polonies arise from the direct amplification of DNA molecules, polony sequencing avoids some of the artifacts introduced by *in vivo* cloning; furthermore, as many polonies can be assayed simultaneously, it is considerably cheaper than conventional Sanger sequencing.

To create polonies, one begins by diluting a library of DNA molecules into a mixture that contains PCR reagents and acrylamide monomer. The mixture is then poured on a microscope slide to form a thin ($\sim 30\mu\text{m}$) gel, and amplification is performed using standard PCR cycling conditions. If one begins with a library of nucleic acids, such that a variable region is flanked by constant regions common to all molecules in the library, a single set of primers complementary to the constant regions can be used to universally amplify a diverse library. Amplification of a dilute mixture of single-template molecules leads to the formation of distinct, spherical polonies. Thus, all molecules within a given polony are amplicons of the same single molecule, but molecules in two distinct polonies are amplicons of different single molecules. Over a million distinguishable polonies, each arising from distinct single molecules, can be formed and visualized on a single microscope slide [2].

Typically, one of the amplification primers includes a 5'-acrydite modification, such that it becomes covalently linked to the gel matrix. Consequently, after PCR, the

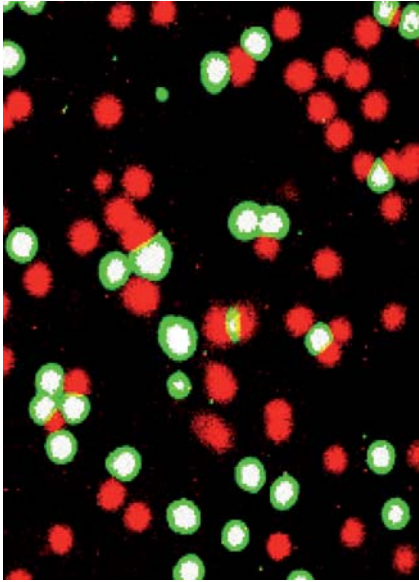


Figure 5.1 Polonies. Each colony of DNA or polony arose from an individual DNA molecule. The polonies have been probed with a fluorescent primer and thus appear red or green.

same strand of every double-stranded amplicon is physically linked to the gel. Exposing the gel to denaturing conditions permits efficient removal of the unattached strand. As every copy of the remaining strand is physically attached to the gel matrix, a variety of biochemical reactions can be performed on the full set of amplified polonies in a highly parallel manner (i.e., additional PCR amplification, single-base extensions, ddNTP extension, hybridization, sequencing by synthesis, etc.) [2–5].

5.2.2

Evolution of Polonies

The polony concept evolved from the original idea of the *in situ* amplification of DNA molecules in an acrylamide gel into a robust approach with the potential to sequence a human genome. During the course of this evolution, there have been many advances; in this section, I will summarize three critical advances.

One of the key advances in this evolution was a complete departure from the *in situ* amplification method. This development was inspired by the BEAMing (beads, emulsion, amplification, and magnetics) method described by Dressman *et al.* [6]. In polony sequencing, BEAMing is used to create the DNA-coated beads, which can be immobilized on a surface. The DNA-coated beads immobilized on a surface can be thought of as polonies (Figure 5.2) and hence the name “polony sequencing” has been retained. The BEAMing approach is thus an alternative to the original polony technology for creating a dense array of distinct colonies of clonal DNA.

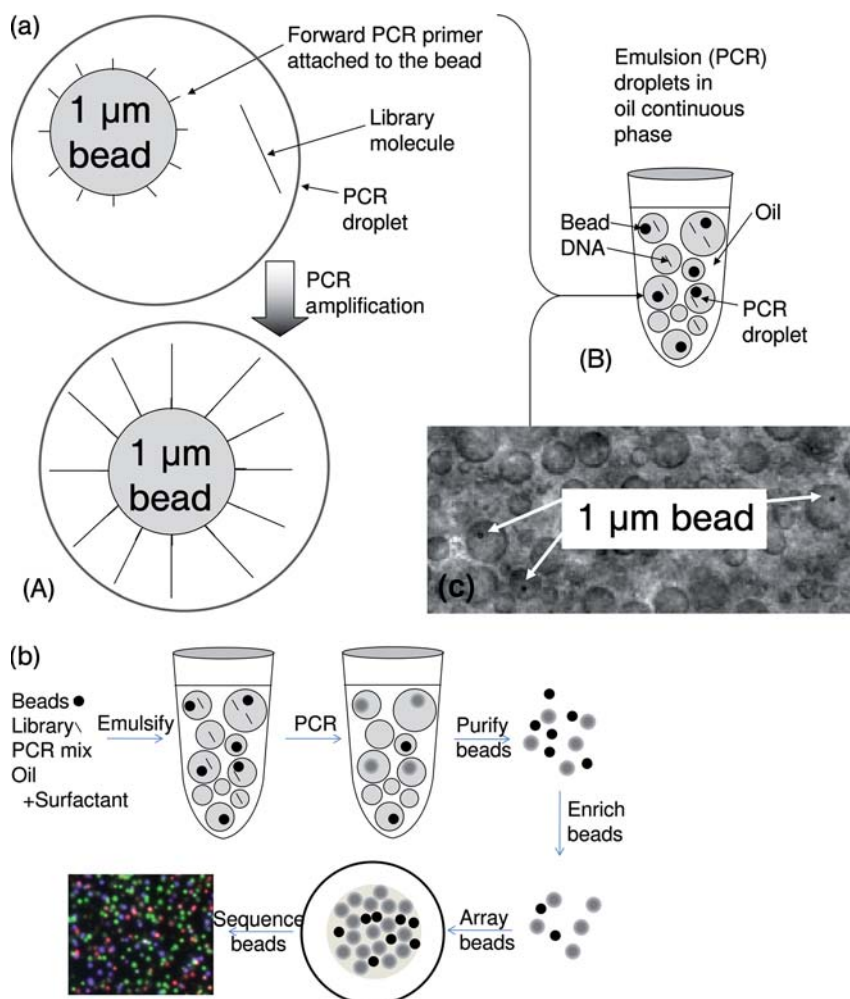


Figure 5.2 Beads, emulsion, amplification, and magnetics or BEAMing. (a) Emulsion droplets are basically femtoliter PCR tubes with individual beads to sequester the clonal PCR amplicons from each droplet. (b) One-micrometer streptavidin-coated magnetic beads are bound to biotinylated forward primers. An aqueous mix containing all the necessary components for PCR plus primer-bound beads and template DNA is mixed with an oil and surfactant to create an emulsion with $\sim 3\text{--}10\mu\text{m}$ droplets of PCR mix.

Each drop is essentially a PCR tube with a volume less than 1 nL. The emulsion is temperature cycled as in a conventional PCR. If a DNA template and a bead are present together in a single aqueous compartment, the bead-bound oligonucleotides act as primers for amplification. Following the PCR, the emulsions are broken, and the beads are purified and enriched. Finally, the beads with DNA are immobilized on a glass coverslip in flow cell and sequenced.

Using BEAMing (also known as emulsion PCR or ePCR) approach for amplification was one of the critical extensions to polony technology for a number of reasons:

- (1) The number of “polonies” that are simultaneously analyzed is greatly increased when using beads [2]. For example, up to 60 million beads can be immobilized on

a surface of $\sim 1 \text{ cm}^2$, which is much greater than the maximum number of polonies that can be traditionally created in the same area. Therefore, BEAMing approach provides a significant increase in the throughput of the polony-based assays.

- (2) The DNA on the beads is more concentrated, hence increasing the signal-to-noise ratio.
- (3) Biochemical reactions are more efficient on the beads compared to DNA immobilized in an acrylamide gel. Furthermore, additional biochemical reactions can be performed on the beads that were not possible with the classical polonies.
- (4) The image analysis of the immobilized beads is easier. The location of each and every bead can be identified and hence only the fluorescent signal is read from each bead location, thus reducing the difficulty associated with identifying the polonies.

The second critical development in the evolution of polony sequencing was the shift to a ligation-based sequencing approach. Many methods for sequencing were attempted over the years, and the method that was most extensively pursued originally was a sequencing-by-synthesis approach [5]. However, the shift to ligation-based sequencing led to an increase in the accuracy of sequencing. The limitation of the ligation-based sequencing method was primarily related to the inability to sequence more than six to seven bases in a single read. To partially overcome this limitation, sequencing libraries for polony sequencing have been constructed in specific manner (see Section 5.3).

The third critical development in polony sequencing was the development of instrumentation to automate the process (see Section 5.3). The shift to a bead-based polony and ligation-based sequencing facilitated the development of polony sequencing automation. Also, there still is room for future improvements in the automation. Currently, the custom-built polony sequencing system is being converted to a completely automated polony sequencing system. This automated system will rival all the commercial, next-generation sequencing systems in terms of capabilities, while costing significantly less. Also, all the software for the polony sequencing system is completely open source and therefore allows the development of customized applications.

5.2.3

Current Applications of the Original Polonies Method

In general, *new* polony sequencing (from beads) has replaced the use of the original polonies based on *in situ* amplified DNA in an acrylamide gel matrix. However, there still remain a few specific applications for original polonies and they are still used. The most common reason for using the original polony method is to study long PCR amplicons. It is currently difficult to generate long amplicons using the BEAMing method, and it is easy to amplify DNA fragments over 1 kb using the *in situ* amplification polony approach. Owing to the amplicon size limitations, original polonies are currently being used to characterize the splicing patterns in transcripts.

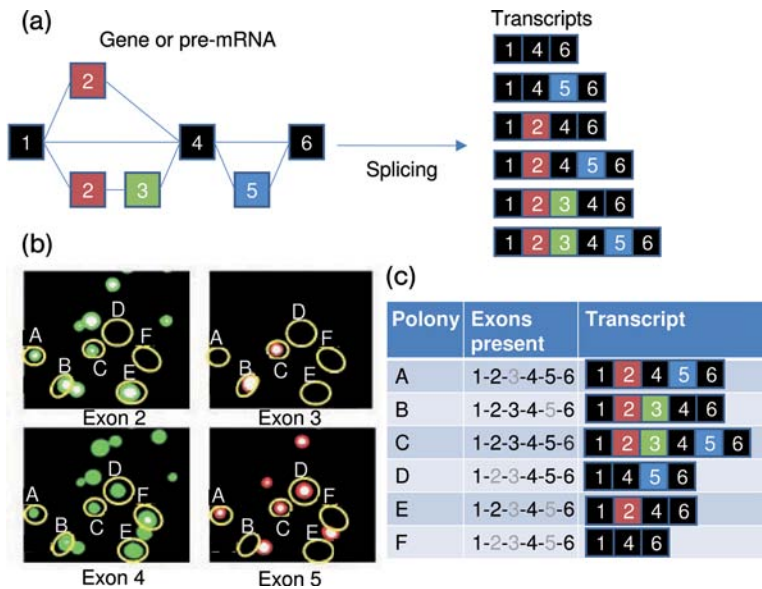


Figure 5.3 Exontyping: measuring alternative splicing by using polony technology. (a) Hypothetical gene with six exons. These six exons can be alternatively spliced to generate six distinct transcripts. (b) Polony gel. Each polony is a colony of amplicons from a single cDNA molecule. Hybridization probes to each exon are

used to identify the exons that are present or absent in each and every cDNA molecule. (c) The transcript that corresponds to each polony is visually shown. When analyzing thousands of polonies, one can identify the expression level of each alternately spliced transcript.

In general, splicing measurements are difficult to make using traditional laboratory tools and methods. Currently the best method for quantifying exon expression is the all-exon arrays that are commercially available. The problem with these arrays is that the exon expression levels are measured independent of the context of the entire transcript. The only method for understanding how the exons are spliced together and quantifying the expression of the potential splice variants in the total mRNA population, other than full cDNA sequencing, is using real-time PCR with primers that span exon junctions. The real-time PCR method is difficult, expensive, and still does not provide information regarding *all* exons present in the transcript. Polony technology is perfect for making this measurement, and this approach has been called digital polony exon profiling [4, 7] or exontyping. This approach is pictorially depicted in Figure 5.3.

5.3 Polony Sequencing

Polony sequencing is a completely modular platform, which is due to the design of the sequencing method and the open source nature of the software and equipment

designs. Each of the steps in polony sequencing is independent, and hence polony sequencing can be used to sequence various libraries from different sources using a variety of methods at each step. In genome sequencing, for example, libraries of paired genome tags are produced; however, libraries with a single tag are frequently sufficient for certain applications, such as gene expression profiling, and for these specific applications single-tag libraries can be constructed. In addition, the BEAMing or emulsion PCR approach is typically used to create a large number of distinct clonal beads, but the approach could be replaced by any other method that produces the same result. Finally, many different biochemical methods for sequencing can be used. For example, a sequencing-by-synthesis strategy based on polymerase extension was originally used, but the approach was abandoned for ligation sequencing. It is possible that certain future sequencing applications could benefit from a sequencing-by-synthesis strategy. Polony sequencing could use any of these methods.

Polony sequencing, as practiced today, involves sequencing six or seven (depending on the sequencing direction) continuous bases from DNA immobilized on 1 μ m beads. Since polony sequencing utilizes the ligation strategy, it is possible to sequence in both directions. Therefore, one is able to sequence a minimum of 13 bases from the immobilized DNA. Often, mate-pair libraries are prepared such that 26 bases of sequence can be obtained from each bead. Since up to 60 million beads can be sequenced in a single run, polony sequencing has the potential to sequence over 1.5 billion bases in a single run. In this section, I will describe the steps involved in sequencing a bacterial genome by using polony sequencing.

In 2005, Shendure and colleagues published the methods for sequencing an entire bacterial genome using polony sequencing. There are essentially five steps involved in this process: (1) constructing a sequencing library; (2) loading the library onto beads by using BEAMing; (3) immobilizing the beads in the sequencing flow cell; (4) sequencing; and (5) data analysis.

5.3.1

Constructing a Sequencing Library

Preparing a sequencing library begins with the isolation of genomic DNA. The genomic DNA is then sheared to ~ 1 kb (or appropriately sized) fragments. The ultimate goal is to sequence the ends of these 1 kb fragments. Since the sequencing reads are very short (six or seven continuous bases) and we are unable to perform BEAMing with DNA fragments much larger than 200 bp, the goal of constructing the sequencing library is to extract 18 bases from each end of these 1 kb fragments and position them as shown in Figure 5.4. To construct the library as shown in Figure 5.4, the ~ 1 kb fragments are isolated and circularized with the circularization primer (red in Figure 5.4). The key features of the circularization primer are the sequencing primer sites and *MmeI* type IIs restriction enzyme sites. The circularized library can be amplified by rolling circle amplification, if needed. Next, the library is digested with *MmeI*, which releases a 70 bp fragment with circularization primer flanked by the 17/18 bp genomic sequences separated in the genome by ~ 1 kb. The 70 bp fragments have two-base 3' overhangs. Primers for the PCR amplification and

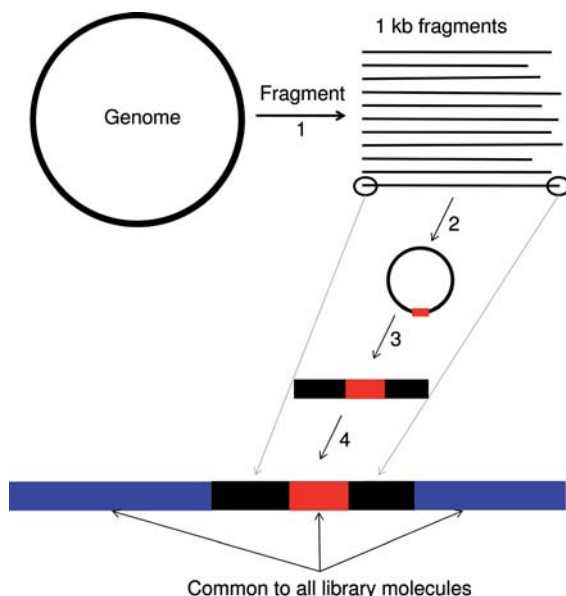


Figure 5.4 The steps involved in preparing a polony sequencing library from bacterial genomic DNA. First, a bacterial genomic DNA is isolated and fragmented to ~1 kb. The 1 kb fragments are circularized with an adapter

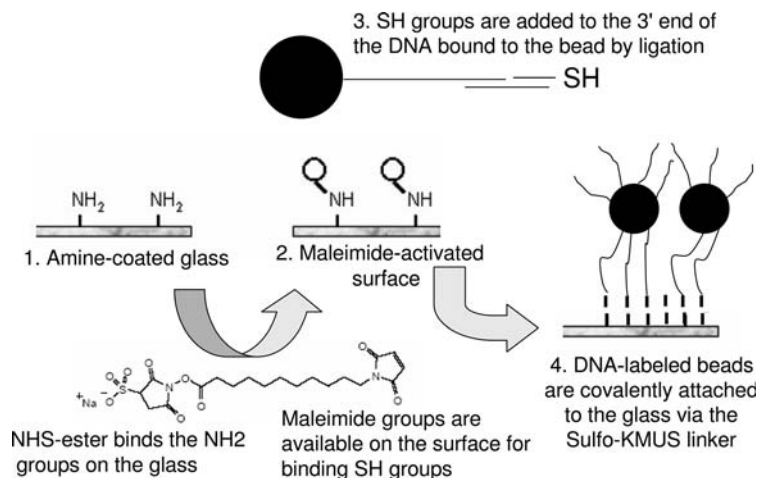
IIs restriction sites. The circularized molecules are digested to generate ~70 bp fragments that contain the circularization primer flanked by the ends of the original 1 kb fragment. Sequencing/PCR primer sites are then ligated onto the 70 bp molecule to generate the final library molecules.

sequencing are then ligated to the 70 bp library fragment using the two-base 3' overhangs. This produces ~134-bp fragments (shown in Figure 5.4). The library of these molecules is then loaded onto beads using the BEAMing approach.

5.3.2

Loading the Library onto Beads Using BEAMing

The sequencing library is next loaded onto 1 μ m beads using the BEAMing method (Figure 5.2). BEAMing is an emulsion PCR method that allows one to load a bead with clonal DNA molecules, and since it is done in an emulsion, millions of distinct clonal beads can be created. The concentration of the library molecules in the emulsion is extremely important. The goal is to get ~20% of the beads loaded with DNA. If a higher percentage of beads is loaded with DNA, too many beads are nonclonal, and too low a percentage of beads loaded with DNA requires additional BEAMing reactions to create a sufficient number of DNA-coated beads for sequencing. Following the BEAMing, the beads can be enriched to maximize the number of DNA-coated beads in the final bead array. The bead enrichment process basically involves selecting beads that have DNA while discarding beads without DNA. DNA-loaded beads are selected by using large polystyrene beads (enrichment beads) with a sequence complementary to



Sulfo-KMUS from Pierce

Figure 5.5 Bead immobilization chemistry. The beads are immobilized in a flow cell by covalently linking the DNA on the beads to an amine-coated surface.

the common 3' end of the DNA-loaded beads. The beads with and without DNA are then separated by density. Ideally, every bead in the array would be clonally coated with DNA. In practice, however, typically 50% of the beads either do not have DNA, are nonclonal, or the signal is too low to generate a high-quality sequence. Therefore, in reality, only 50% of the beads in the array will provide sequence in a typical sequencing run.

5.3.3

Immobilizing the Beads in the Sequencing Flow Cell

The beads are immobilized in the flow cell for sequencing by covalently attaching them to a glass surface. The immobilization chemistry is illustrated in Figure 5.5. Briefly, an oligonucleotide with a 3'-SH group is ligated to the 3' end of the DNA on the beads (placing a 3' "cap" on the DNA). Since the 3' sequence of the DNA on the beads is common to all molecules, a bridging oligo can be designed to allow the ligation of an oligo to the 3' end of the DNA coating the beads. The 3' modification on the library molecules serves two important roles. First, it blocks the 3' end of the library and prevents ligation of the sequencing nonamers onto the 3' end during the sequencing by ligation steps. Nonspecific ligation to the 3' end of the library molecules is suspected to be a major contributor to background signal buildup on the beads during the sequencing. The background signal increases after sequencing with phosphorylated nonamers (sequencing in the 3' to 5' direction; Figure 5.6). Second, the 3' cap provides a functional group for chemical cross-linking to a glass surface (surface inside the flow cell). As already mentioned, up to 60 million beads can be immobilized on a glass surface in an area of $\sim 1 \text{ cm}^2$.

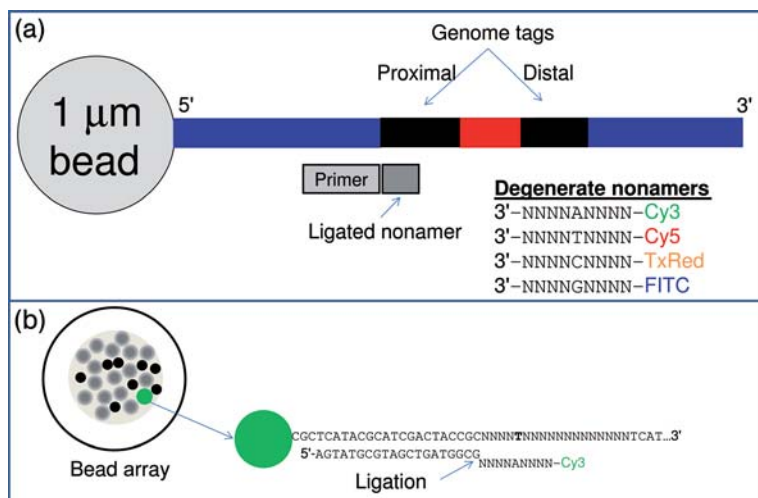


Figure 5.6 Ligation sequencing of the DNA-coated beads in the array. (a) Library molecule on the bead. In this figure, the library molecule contains two tags, such as the mate pair tags that are used for genome sequencing. The tags are sequenced by ligation sequencing in both

directions (toward the bead “+ direction” and away from the bead “– direction”). A different set of nonamers is required to sequence each position in each direction. (b) Example sequencing of a single position.

5.3.4

Sequencing

Initially, polony sequencing was performed using a sequencing-by-synthesis strategy [5], and given the versatility of polony sequencing, one could easily perform sequencing-by-synthesis within the flow cell by simply modifying the software that runs the fluidics system. However, sequencing by ligation is the commonly used method in polony sequencing, and the switch to sequencing by ligation was one of the major advances that made the technology a feasible approach for large-scale sequencing projects.

In sequencing by ligation, the DNA on the beads is sequenced by the specificity of the ligation reaction. The method was inspired by the massively parallel signature sequencing (MPSS) strategy developed by Brenner and coworkers [8]. The DNA tag is sequenced by repeated rounds of denaturing (making the bead DNA single stranded), annealing an anchor (sequencing) primer, and ligation of a set of four fluor-labeled degenerate nonamers (9-mers) to the primer. The degenerate nonamers are degenerated at every position except for the base that is being sequenced. The position for the base that is being sequenced is fixed and a fluor is attached to the nonamer that correlates with the fixed position. For example, 5'-Cy3-NNNNNANNNN is one of the four nonamers used to sequence the fifth position away (in the 3' direction) from the ligation position. Following the ligation step, the beads are imaged with a microscope so that the nonamer that ligated to each and every bead can

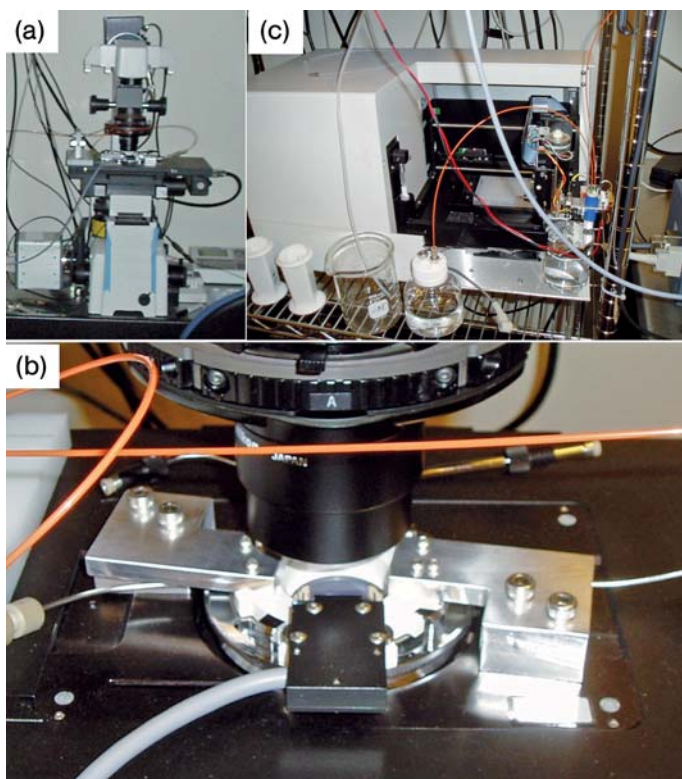


Figure 5.7 Polony sequencing system. (a) Inverted microscope for imaging the bead array. (b) Flow cell containing the bead array mounted on the automated stage of the microscope. (c) Autosampler that injects the biochemical reactions into the flow cell at specific times. The entire system is computer controlled.

be determined visually. Upon visual inspection, the base at a given position on all beads can be determined by the fluorescent signal.

To realistically perform polony sequencing on a large scale, an automated system is required (Figure 5.7). Therefore, all the steps for polony sequencing have been automated by using a computer-controlled automated stage with a flow cell containing the bead array that is mounted on an inverted fluorescent microscope. The flow cell is automatically loaded with all the sequencing reagents by computer-controlled syringe pumps connected to a 96-well plate autosampler. In addition, there is a 1-megapixel camera attached to the microscope capable of collecting 30 images per second that automatically collects images of the entire polony array after completion of the ligation step (Figure 5.7). The part list and design for the complete polony sequencing system are freely available online. Furthermore, all the source code for controlling the fluidics, imaging, and data processing is also available online. Future polony sequencing systems will not require the engineering and software expertise of the original design. There is a completely automated polony sequencing system

available from Danaher Motion, Salem, NH, USA. The Danaher Motion system features increased automation and has the capability of running multiple bead arrays simultaneously. Thus, the throughput of polony sequencing will greatly increase relative to the original custom sequencers built with standard components.

5.3.5

Data Analysis

Polony sequencing produces a large amount of raw data. For example, approximately 2000 frames are required to cover the entire array of beads, and for each frame four images are acquired (one for each base). Hence, there are ~ 8000 images for each position, and if we are sequencing 26 bases, then over 200 000 images are required for the complete sequencing run (>200 gigabases). For each base that is sequenced, every bead on the entire array must be found and matched with exactly the same bead for every other position. When each bead is found, the fluorescence intensity in four channels is read (Figure 5.8). This information is then stored and once all bases are sequenced, the entire sequence for each bead is saved to a text file.

Processing all four images for each of 2000 frames is computationally very intense. To process the images, all four images for a frame need to be aligned to a “base” image where all the beads are identified. The fluorescence intensity in each of the four images is then read from each bead in the frame, and the base is called as being the base with the highest intensity fluor signal. A quality score is also calculated for each base, and the quality score is defined as the Euclidean distance from the perfect base call (which would be high signal from one fluor and all others zero; Figure 5.8). It

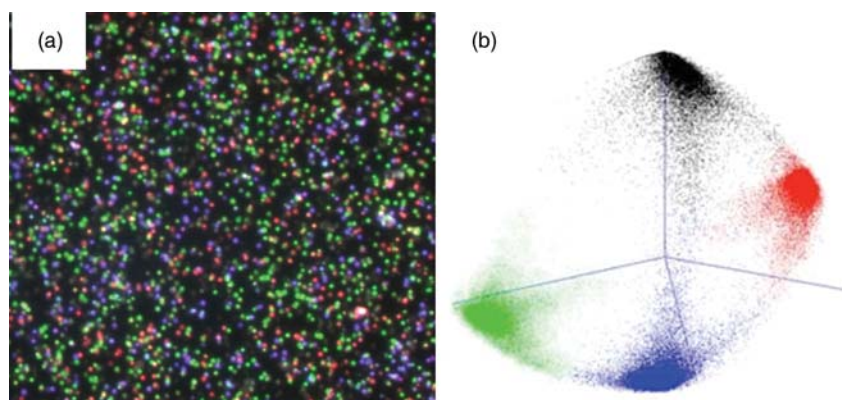


Figure 5.8 (a) Raw sequencing image ($\sim 1/4000$ th of the entire array). Each different color indicates a base that is read at each position. Only three colors (and 75% of the beads) are shown because the software would only allow us to overlay three images simultaneously. (b) The images were processed to generate base calls for

each bead. The intensity of each of the four color channels is plotted in a 4D tetrahedral, as shown. The quality of the base call is defined as the sum of the distances from point to the centroid of the cluster for each sequenced base (only one base shown).

takes approximately 1 h to process all the images and call a single base for every bead on an array using a standard Windows PC. There is little motivation to improve the speed performance of this step because it takes approximately 1.5 h to perform the biochemical sequencing reactions on the array; therefore, all the processing is performed during the biochemical reaction (sequencing by ligation) performed for the next base. However, if increased speed is required, the processing could be simply parallelized by having each multiple processor process different sets of frames.

Once the sequencing and analysis of the images are complete, a file that contains the DNA sequence on each bead is generated. Next, this file of individual reads is processed to generate the sequence for the entire genome of interest. Currently, polony sequencing works very well for finding single-nucleotide differences between the sequenced genome and a reference genome. The single-nucleotide differences are found by mapping the raw sequence back to the reference genome. To improve the sequencing accuracy, low-quality positions in the reads are neglected in the mapping and thus, mismatches that are identified in low-quality positions in the reads are neglected. Then, all the reads (usually much greater than $10\times$ coverage) are mapped to the reference genome and positions where nucleotide changes are consistently found are identified. In the end, all the nucleotide differences are written to a file, and polony sequencing can generate assembled genome sequence with an overall final accuracy of better than one error per million bases.

5.4 Applications

Since polony sequencing is a completely open source sequencing framework, there are many potential applications, and the user has the ability to customize the system to any desired specification. In this section, I will discuss two general application areas where polony sequencing is likely to have an impact: human genome sequencing and transcript profiling.

5.4.1 Human Genome Sequencing

Polony sequencing is a technology that obtains very short reads. Therefore, the *de novo* sequencing of the human genome is likely impossible. However, with the availability of several human genome sequences, future human genome sequencing projects will be resequencing efforts aimed at identifying the differences between the genome of interest and the available sequences.

5.4.1.1 Requirements of an Ultrahigh-Throughput Sequencing Technology

When considering an alternative approach for resequencing the human genome, one must consider the accuracy requirements of the project and determine that the approach is appropriate. The minimum accuracy requirement is error rate of 1 per

10 000 bases in the final assembled sequence. Polony sequencing can provide genome sequence with an error rate much better than this; for example, Shendure *et al.* sequenced an *Escherichia coli* genome with less than one error per million bases.

The final assembled accuracy of the sequence is influenced by two factors, the raw error rate and the coverage. Polony sequencing has a raw error rate of 99.7% or 3 errors per 1000 bases, which is on par with Sanger sequencing [9]; therefore, if one obtains $3 \times$ coverage of each base the resulting sequence will have an assembled error rate of $1/100\,000$ bp. To ensure a minimum of $3 \times$ coverage of more than 95% of the human genome, approximately 40 billion raw bases must be sequenced ($\sim 7 \times$ coverage). This goal can be met by polony sequencing. The automated polony sequencing system is capable of sequencing over 56 billion bases. Realistically, typically 50% of the sequencing reads will pass the quality control requirements; therefore, polony sequencing would likely generate 28 billion bases in 2.5 days. Hence, polony sequencing could potentially sequence the human genome in less than 1 week.

5.4.2

Challenges of Sequencing the Human Genome with Short Reads

1. *SNP identification.* It is difficult to resequence the human genome with short reads because there are a substantial number of recently duplicated sequences in the genome [10]. Shendure and colleagues estimate that sequencing reads greater than 200 bases will be required for more than 99% of the sequences to uniquely match a sequence in the human genome [11]. With 13-base mate-paired reads, polony sequencing can cover 83–85% of the human genome (Figure 5.9) and with modest improvements in the read length at least 95% coverage.

Figure 5.9 Whole genome simulations for 381×10^6 reads of mate-paired 13 bp tags separated by 1000 ± 300 bp. The columns in the figure indicate the chromosome where the tags were generated and the rows indicate the chromosome for which the tag matched. The highlighted cells along the diagonal indicate the percentage of tags from a chromosome that uniquely mapped correctly to the same chromosome (i.e., the result if the sequencing was performed on flow-sorted chromosomes). All other rows indicate the percentage of tags that incorrectly matched a different chromosome. The bottom row (“Good Tags”) indicates the final percentage of tags that correctly matched the given chromosome. The total at the bottom is the final percentage of correctly mapped tags. The “Good Tags” number is smaller than the number on the diagonal because a correctly

mapped tag to the correct chromosome may still be a “bad” tag if it incorrectly matches another chromosome. The total indicates the total percentage of tags that correctly matched all chromosomes. Therefore, with 13 bp paired reads, we anticipate correctly mapping 80% of the tags. Based on an extrapolation of chromosome simulations, we anticipate this will provide $\sim 85\%$ sequence coverage. Additionally, we can see that we will obtain very good coverage of chromosome 13 (83% tags matching correctly) and very poor coverage of the Y-chromosome (36% of tags matching correctly). The numbers are improved with longer reads (data not shown): 14 bp reads = 86% total tags matching correctly; 15 bp reads = 89% total tags matching; 16 bp reads = 90% total tags matching.

2. *Defining haplotypes.* Sequencing diploid genomes with short reads has its drawbacks, namely, it is not possible to identify haplotype blocks without prohibitively high sequencing coverage.
3. *Chromosomal rearrangements, large insertions, deletions, and amplifications.* Detecting chromosome alterations is generally important; however, it is particularly important for sequencing tumor genomes. It can be difficult to identify chromosomal abnormalities and polymorphisms, and it is more challenging than identifying single-nucleotide polymorphisms (SNPs) via short reads. Since polony sequencing allows the sequencing of mate pairs, rearrangements and large insertions and deletions can be detected.
4. *Small insertion and deletions.* The information on small insertions and deletions is contained in the raw data. However, the software to extract this information is not currently available.

5.4.2.1 Chromosome Sequencing

Difficulties associated with sequencing the human genome can be reduced by sequencing isolated chromosomes (Figure 5.9). Approximately, a 10% higher mapping rate for tags can be obtained when mapping to a single chromosome rather than the entire genome. Ideally, shotgun sequencing the entire genome would be easier since it would not require the additional step of flow sorting the chromosomes. However, the effort associated with sorting the chromosomes could significantly increase the coverage and make human genome sequencing with very short reads feasible.

5.4.2.2 Exon Sequencing

As an alternative to whole genome sequencing, one may focus attention on sequencing only the coding regions of the genome, since the coding region of the genome represents only a small fraction of the 3×10^9 bp haploid genome. Sjoblom *et al.* [12] defined the consensus coding sequences that represent the most well-annotated genes in the genome, which corresponds to $\sim 21 \times 10^6$ bp. Sequencing 21×10^6 bp is relatively simple for polony sequencing; however, the new challenge is preparing a sequencing library that represents the small fraction of the human genome. Church and coworkers have recently developed a multiplex PCR approach directly designed to construct polony sequencing libraries that represent the coding region of the genome. This will likely be a powerful tool for population-based studies, because the reduced sequencing sample size will allow a large number of people to be sequenced.

5.4.2.3 Impact on Medicine

The availability of genome sequences has forever changed the way biomedical research is performed. The ability to generate genome sequences more rapidly will undoubtedly have many medical implications. Currently, a personal genome project may be of minimal medical value; however, once many genomes are available, we will have a very powerful tool for uncovering the associations between the genotype and the phenotype. The prospect of having many genomes and personal genome projects relies on inexpensive sequencing technologies, such as polony sequencing. In addition, sequencing of specific genomes related to disease, that is, sequencing a

cancer genome, will be of particular value by identifying the specific somatic mutations that have occurred during neoplastic transformation.

5.4.3

Transcript Profiling

The short reads of polony sequencing are ideally suited for serial analysis of gene expression (SAGE) library sequencing. SAGE libraries consist of tags from mRNA that can be used to quantify the expression level of all the genes in a cell. There are several important advantages of SAGE for studying gene expression over microarray-based measurements: (1) SAGE profiles all expressed genes and is not subject to investigator bias in arraying only known genes. (2) SAGE is more quantitative compared to array-based methods. (3) SAGE has a larger dynamic range in measuring gene expression. (4) SAGE is not subject to cross-hybridization errors. The main disadvantage of SAGE is that it is very labor intensive and expensive to sequence the SAGE tags by traditional methods. Polony sequencing is ideal for sequencing SAGE libraries as the inherent short reads are ideal for SAGE libraries, and the disadvantages of traditional SAGE are eliminated because the SAGE library preparation for polony sequencing is simpler and the sequencing is extremely cheap and millions of tags can be sequenced simultaneously. Finally, if multiple SAGE libraries are “barcoded” with DNA tags and sequenced simultaneously, the sequencing costs can be dropped to ~\$10 per library.

5.4.3.1 Polony SAGE

Polony sequencing of a SAGE library requires the preparation of a SAGE library that is different from the classical concatemer of ditags typically used in SAGE. Actually, the preparation of a SAGE library for polony sequencing is relatively straightforward [13]. The essential steps for preparing a polony SAGE library are described in Figure 5.10. The final SAGE library for polony sequencing is simply an mRNA tag adjacent to an anchoring enzyme site (commonly *Nla*III) flanked by two sequencing/PCR primer sites. The library is loaded onto beads and sequenced identically to the genome libraries described above. The only difference is that only a single tag (unitag) is present in the library molecules and hence the library is only ~100 bp. The unitag can be sequenced from both directions; therefore, 13 bases can be sequenced from the tag.

5.4.3.2 Transcript Characterization with Polony SAGE

Full cDNA sequencing is of great value for characterizing the transcripts, that is, identifying the transcriptional start site, transcriptional stop site, and all exons present in the mRNA. However, full transcript sequencing on a large scale is a formidable task. Furthermore, this measurement cannot be made with array-based technologies, and exact start and stop sites can only be reliably identified with a sequencing-based approach. Alternatively, sequencing mate-paired transcriptional start and stop sites provides valuable information for finding new genes, understanding the regulation of known genes, and studying the signals that cause shifts in the transcriptional start and stop sites.

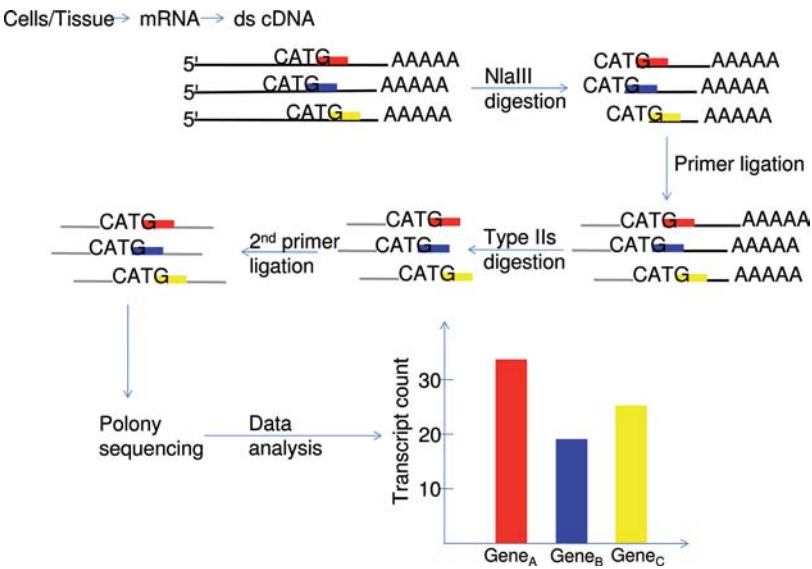
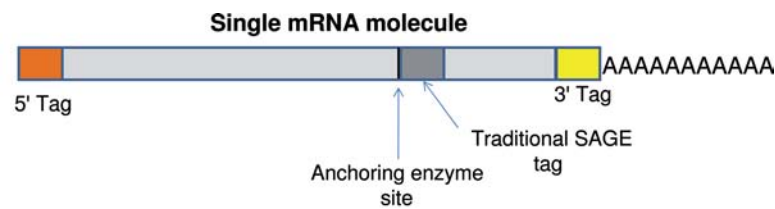


Figure 5.10 Polony SAGE. Tags can be extracted from cDNA to generate a polony SAGE library for gene expression profiling. The steps required to prepare a polony SAGE library are illustrated.



Library molecule	Description
	mRNA tag adjacent to anchoring enzyme site
	mRNA tag from the 5' end of the transcript
	mRNA tag from the 3' end of the transcript
	Paired 5' and 3' mRNA tags

Figure 5.11 Modified SAGE approaches for characterizing the transcriptome. Tags can be extracted from various locations from the cDNA molecules. Traditional SAGE-extracted tags adjacent to an NlaIII site. Tags from the 5' or 3' end can be used to map the start and stop sites for transcription and can provide valuable biological insight.

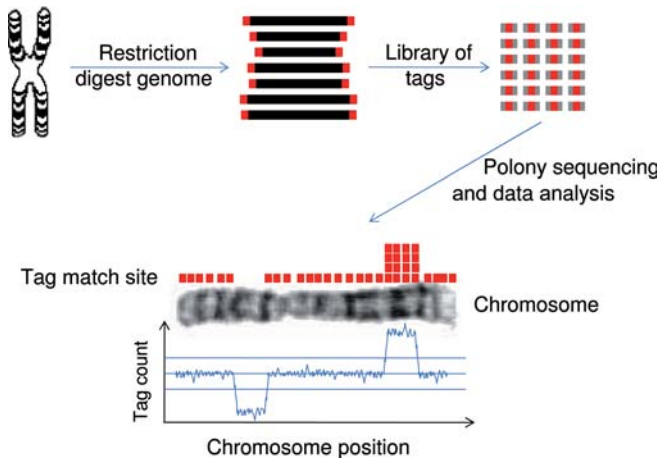


Figure 5.12 Polony karyotyping. Tags can be extracted at specific location across the entire human genome. For example, tags can be isolated next to a restriction site. The tags can then be sequenced by polony sequencing, and the tag density across the entire genome can be plotted. Genomic regions that are amplified or deleted can easily be identified from such a study, thus indicating hypothetical oncogenes and tumor suppressor genes.

In general, one could perform polony SAGE to identify various regions of the transcript (Figure 5.11). The library type selection depends on the goals of the study. For example, for gene expression measurements, the traditional SAGE tag library is the best choice. One of the advantages of polony SAGE is that one could consider constructing multiple libraries with various anchoring enzymes to increase the coverage of the transcriptome. If one is interested in finding and identifying new genes and transcripts, the mate-paired library with the beginning and the end of transcription is the appropriate library. However, one should recognize that libraries that contain the 5' end of the transcript would likely be biased due to difficulties associated with obtaining full-length cDNA.

5.4.3.3 Digital Karyotyping

There are a number of SAGE-like approaches that can be significantly improved with next-generation sequencing technologies. A good example is the approach known as digital karyotyping, which was developed to identify chromosomal amplifications and deletions by sequencing genome tags next to restriction enzyme site and mapping the tag count across the entire genome (Figure 5.12).

5.5 Conclusions

The availability of genome sequences has forever changed the way biomedical research is performed. The ability to cheaply generate genome sequences

(for pathogens and humans) very rapidly will undoubtedly have many medical implications. Ultimately, the value of next-generation sequencing technologies will be in sequencing a large number of samples. For example, the ability to sequence hundreds of pathogenic bacteria responsible for various diseases will provide important information toward understanding the evolution of infectious diseases and emergence of antibiotic resistant strains. The technology is rapidly evolving and will soon allow large-scale sequencing projects to study hundreds to thousands of human genomes. Initially, these projects will likely be devoted to exon sequencing and soon evolve toward whole human genome sequencing. Currently, having a personal genome project may be of minimal medical value; however, once many genomes are available, we will have a very powerful tool for uncovering the associations between the genotype and the phenotype.

References

- 1 Collins, F.S., Morgan, M. and Patrinos, A. (2003) *Science*, **300** (5617), 286.
- 2 Mitra, R.D. and Church, G.M. (1999) *Nucleic Acids Research*, **27** (24), e34.
- 3 Mikkilineni, V., Mitra, R.D., DiTonno, J.R. *et al.* (2004) *Biotechnology and Bioengineering*, **86** (2), 117; Merritt, J., DiTonno, J.R., Mitra, R.D. *et al.* (2003) *Nucleic Acids Research*, **31** (15), e84; Mitra, R.D., Butty, V.L., Shendure, J. *et al.* (2003) *Proceedings of the National Academy of Sciences of the United States of America*, **100** (10), 5926; Butz, J., Wickstrom, E. and Edwards, J.S. (2003) *BMC Biotechnology*, **3** (1), 11; Butz, J., Yan, H., Mikkilineni, V. *et al.* (2004) *BMC Genetics*, **5** (3).
- 4 Zhu, J., Shendure, J., Mitra, R.D. *et al.* (2003) *Science*, **301** (5634), 836.
- 5 Mitra, R.D., Shendure, J., Olejnik, J. *et al.* (2003) *Analytical Biochemistry*, **320** (1), 55.
- 6 Dressman, D., Yan, H., Traverso, G. *et al.* (2003) *Proceedings of the National Academy of Sciences of the United States of America*, **100** (15), 8817.
- 7 Butz, J., Goodwin, K. and Edwards, J.S. (2004) *Biotechnology Progress*, **20** (6), 1836.
- 8 Brenner, S., Johnson, M., Bridgham, J. *et al.* (2000) *Nature Biotechnology*, **18** (6), 630.
- 9 Shendure, J., Porreca, G.J., Reppas, N.B. *et al.* (2005) *Science*, **309** (5741), 1728.
- 10 Bailey, J.A., Gu, Z., Clark, R.A. *et al.* (2002) *Science*, **297** (5583), 1003.
- 11 Shendure, J., Mitra, R., Varma, C. *et al.* (2004) *Nature Reviews Genetics*, **5**, 335.
- 12 Sjoblom, T., Jones, S., Wood, L.D. *et al.* (2006) *Science*, **314** (5797), 268.
- 13 Kim, J.B., Porreca, G.J., Song, L. *et al.* (2003) *Science*, **316** (5830), 1481.

Part Three

The Bottleneck: Sequence Data Analysis

6

Next-Generation Sequence Data Analysis

Leonard N. Bloksberg

6.1

Why Next-Generation Sequence Analysis is Different?

The idea of sequence analysis is familiar to most of us, but the challenges of Next-Generation Sequencing (NGS) are forcing some new thinking and new strategies. This is an active area of research, and the intention of this chapter is to review issues and strategies applied to DNA sequence analysis by NGS. The key difference of NGS data can be summed up in a word: Extreme.

A single NGS run dumps about 100 million nt onto your hard disk drive (HDD) (roughly 400 k reads at 250 nt for 454 or 4000 k reads at 25 nt for Solexa or SOLiD). You will need 30 runs or 120 million reads to cover the human genome 1×. The human genome project required 15× coverage but the current opinion is that 30× coverage will be required with the shorter NGS reads, so you will need 900 runs or 3.6 billion reads or 90 billion nt or 150 GB of FASTA format data on your HDD to resequence a single person. A personalized medicine database with assembled genomes for just 10% of United States population would hold 10⁵ TB of FASTA format data to be searched. With data sets of this size, we encounter significant I/O problems, and problems with file and directory size limits in current operating systems. Simple searches can take years, and many tasks are just not possible with traditional tools.

NGS produces short reads that are frequently searched against large chromosomes, each raising particular problems. Every entity (read) requires a header to be established in memory. Because NGS reads are so small and so numerous, massive resources can be tied up simply establishing these headers. Conversely, a single large sequence cannot be segmented without complex data handling, and all interactions over the entire length of that sequence must be held in memory as a group. Optimizing for one of these extremes often makes significant compromises for the other.

NGS reads often include errors. Methods that are sensitive enough to match all or most reads are usually computationally too intensive, and also find too many false positives, while methods that are discrete and efficient enough to complete in a relevant time, and find few false positives, unfortunately find too many false

negatives. With 3.6 billion reads, a small rate of false positives can overwhelm any real matches. Conversely, each 1% gain in false negatives recovered equates to 1 000 000 nt of data recovered, or about 10 full machine runs fewer per person sequenced. BLAST often loses near-perfect matches of short reads due to the way scores are calculated. SSAHA-based methods build arrays with a sampling offset (step) equal to word size, but with short reads this results in insufficient sampling in the array to detect many perfect matches. SLIM Search and some recent SSAHA implementations allow arrays to be built with step = 1, which seems to be essential for NGS data.

Many parts of the analysis could be done more efficiently if read lengths were consistent. Some researchers are trying to adjust samples to constant read lengths by culling outliers, filling in shorter reads, assuming the longest sequence as the universal length, or grouping the reads into discrete size groups.

Typical NGS analysis is concerned more with “correct” matches (mapping reads or annotation tagging) and less with distant relationships. Because NGS often surveys many closely related species, or many individuals of a single species, discriminating close relationships becomes more important than distant relationships. Repeat regions are conceptually no different, but shorter reads mean the number of repeats that require secondary data (e.g., paired reads) to resolve increases dramatically. A typical NGS analysis pipeline might involve filtering, base calling, and sequence alignment followed by assembly or other downstream tasks. The filtering and base calling is usually done by the manufacturer’s software (e.g., PyroBayes [1] for 454 and Bustard for Solexa). While some people are working on *de novo* assemblers for NGS [2–6], many focus on remapping reads.

Why NGS data analysis is different?

- Extremely large data sets create problems for I/O and file size limits,
- Extremely short reads create data handling problems and scoring anomalies,
- Extremely large chromosomes create data handling problems and scoring anomalies,
- Error rates create problems in sensitivity and data loss,
- Variable read lengths create data handling problems.

A variety of strategies are discussed to manage these challenges.

6.2

Strategies for Sequence Searching

The basic idea of sequence searching is to find matches by strength of similarity. Unfortunately, computers only work in 0 or 1 (match or not), and similarity is not allowed. The typical strategy is to chop the sequence into sections and look for fragments of identity. The Dynamic Programming Matrix (DPM) used by Smith and Waterman [7] and Needleman and Wunsch [8] breaks the sequence into the smallest possible units (individual nucleotides) and plots the highest scoring diagonal through the matrix of short exact matches. This method remains the gold standard;

unfortunately, it displays $O(m)$ & $O(m \times n)^{1)}$ space and time complexity and is not possible for tasks of any size (a human by mouse search would require 10^9 GB of RAM).

The great innovation of BLAST [9, 10] was to look for larger chunks, and search in a one-dimensional array. As a result, BLAST displays $O(m)$ & $O(n)^{1)}$ space and time complexity. Although not as precise as the DPM, it is close enough, and (more important) it is fast enough, to be practical (at least until NGS). BLAST uses exact matches of 11 nt (default) to anchor a local DPM and all final data are generated by the DPM. While the BLAST method has many advantages, it also has limitations, some of which are critical for NGS.

Several authors have proposed new strategies for building, scanning, and processing data from the arrays of short exact matches, as well as strategies for eliminating the need for the DPM. Computer science teaches that the laws of physics constrain how you can build an array. Programming languages, however, contain a variety of methods for working with data in array structures.²⁾ It is possible to create novel permutations by combining aspects of methods as well as by clever data handling.

One of the most significant of the new methods is SSAHA [11], which eliminates the DPM entirely and displays $O(\alpha^k + n/k)$ & $O(m)^{1)}$ space and time complexity resulting in a significant performance improvement over BLAST. Most of the other offerings appear to be permutations of SSAHA [12–20], limited to the performance of SSAHA but with some improvements in sensitivity or other downstream data processing, although a few are permutations of BLAST [21, 22]. A key limitation of SSAHA-based methods is that they are very restrictive, physically limited to word sizes 7–14. One of the few methods to implement an array structure different from BLAST or SSAHA is SLIM Search [23], which displays $O(n/k)$ & $O(m \log(n/k))^{1)}$ space and time complexity. This provides dramatic performance gains over both BLAST- and SSAHA-based methods, particularly as the size of the data sets increases, and the nature of the underlying array structure seems to allow greater flexibility for novel data handling. In our hands, SLIM Search mapped back 99.99% of 1.48 million 454 reads to the yeast genome in 11 min, 96% of 4.1 million Solexa reads to a 55 kb template in 30 min, and 99.99% of 3.4 million SOLiD reads to a 2 MB template in 40 min, on a dual processor Linux work station with 4 GB RAM.

1) Computational complexity is measured with Big-O notation. We use the following symbols to describe the factors that contribute to complexity in sequence analysis: m , length of the query sequence (or query data set); n , length of the subject sequence (or subject data set); k , word length (window size or k -tuple); α , alphabet length (e.g., 4 for DNA); w , number of nonoverlapping words, or k -tuples (n/k); s , number of sequence entries; l , average length of sequence entries.

2) Hash tables, suffix arrays, and a few other methods are specific types of array structures. The literature is quite complex with some authors insisting on very strict definitions, and others using the terms more loosely. We have tried to avoid this issue and all methods are included under the more generic term “array” in this discussion.

6.3

What is a “Hit,” and Why it Matters for NGS?

For simple gene analysis the definition of a hit does not seem important, but for NGS data it is critical. The literature confounds the problem by using terms like HSP and HSSP interchangeably to refer to several different relationships. We have found it necessary to create a vocabulary with six distinct types of hits.

6.3.1

Word Hit

An individual k -mer match between two sequence sections, the smallest possible unit of a “Hit” in BLAST, SSAHA, or SLIM Search and a single nucleotide match in a DPM. Word Hits are primarily managed internally for data analysis and rarely reported to users.

6.3.2

Segment Hit

A region of sequence similarity containing one or more contiguous Word Hits (expanded with a DPM in BLAST to become a local alignment). This is the only kind of hit that BLAST can report.

6.3.3

SeqID Hit or Gene Hit

A match between two sequence entities, which can be identified by an ID, containing one or more Segment Hits. This is the concept of a “Hit” that most biologists relate to.

6.3.4

Region Hit

Special case used to map coding regions to a chromosome (or fragment) when searching a chromosome against SwissProt or UniRef. A “Hit” is defined by $[\text{SeqIDa-Position}] + [\text{SeqID1}]$ (where a is the first in a small set of large sequences and 1 is the first in a large set of small sequences). Typically, the user wants to limit output to the top n hits per region on the chromosome, but the user also wants at least n hits in each region (if available). The $-k$ function of BLAST attempts to achieve this, and the Region-Hits utility of SLIM Search provides a good example. The hemoglobin gene may find a match in five places on a chromosome, and a biologist will need to see all of them, but in each place where it hits, the biologist may only need to see the top three proteins that hit there. A Region Hit is required to resolve this.

6.3.5

Mapped Hit

Special case used to map reads to a template typically when mapping raw sequence reads back onto a chromosome or reference genome. A “Hit” is defined by $[\text{SeqID1}] + [\text{SeqIDa-Position}]$ (where 1 is the first in a large set of small sequences and a is the first in a small set of large sequences). Typically, the user wants to limit output to the top n locations on the template where each read can be mapped, but the user also wants all available hits at a location. A single sequence read may map to five places on a chromosome, but the biologists only needs to see the “correct” place (best hit); however, there may be 17 reads that all map to that same location, and the biologists needs all of them. A Mapped Hit is required to resolve this.

6.3.6

Synteny Hit

Special case of Segment Hits where the number of Sequence IDs is very small and the number of segments is very large; used when searching whole genomes against each other.

BLAST avoids this discussion by reporting local alignment scores (segment hits) as if they are SeqID Hit values. While the highest scoring local alignment in a gene pair is often a reasonable estimate of the similarity of the entire sequence pair, the number of important exceptions becomes critical with the large number of comparisons dealt with in NGS. As a result, using BLAST scores frequently leads to misleading and incorrect conclusions about sequence relationships because the scores are reported for the wrong entity [24].

The Region, Mapped, and Synteny hits are somewhat specific to NGS, and are not dealt with adequately by any available tool yet. The loose handling of the concept of hits is not adequate for the needs of NGS, and the data handling methods provided by older tools can lead to problems. SLIM Search has started to provide some utilities to achieve the required data handling, but there is not yet any integrated solution available.

6.4

Scoring: Why it is Different for NGS?

Fundamentally, a score is just a relative value of quality. In the case of NGS, scores are required to rank and sort the massive quantity of data. The values must be comparable both within and between searches, must be simple enough to be calculated in a relevant time but specific enough to resolve subtle differences, and (ideally) should have some biological significance. Finally, the score should reflect length, substitutions, and indels in a relevant manner.

BLAST scores [25] rely on a DPM, and that is not practical for NGS data sets. Even if you could run the analysis, there are problems with BLAST scores for NGS. Because

BLAST scores are only reported for Segment Hits, scoring anomalies can result, for example, when a pair of 5 kb genes (A and B) share a short 50 nt region of 100% identity, as compared to another 5 kb pair of genes (C and D) that share a total of 60% identity along their entire 5 kb length. The short-read lengths and long chromosomes used result in scoring anomalies, so good matches are often lost below thresholds, and scores cannot always be compared between searches. Although the *p*-value, BLAST score, and *E*-values are useful for simple analysis, they violate many technical assumptions, some of which become critical for NGS. Despite this, the BLAST scores have proved quite useful, and they provide the best handling of indels.

Currently, most strategies for scoring NGS matches focus on some estimate of the region of identity such as percentage identity or the absolute number of mismatches when read lengths are less than 100 nt (and % id is not valid). This can be determined without a DPM by mapping all possible *k*-mers and indexing back to a table of relationships. This method encounters challenges with heterogeneous length sequences, SSAHA-based methods cannot map *k*-mers larger than 14 nt, and complexity of relationships becomes prohibitive for more than three mismatches. Some methods report the number of *k*-mer matches with small word sizes as a simple estimate [26, 27]. This method is fast, but may lack the resolution required when several hits are similar [28].

Unfortunately, none of the newer methods adequately deals with indels. Simply scoring an indel the same as a mismatch is computationally challenging. In a 25 nt read with an indel at position 15, most methods will report 10 nt of mismatch, not 1. The method of counting the number of short *k*-mers in the hit is one way to solve this problem, but only if matches on both sides of the indel can be combined into a Mapped Hit. Other methods are currently being researched.

6.5

Strategies for NGS Sequence Analysis

Because the data sets are so large, it is important to reduce file sizes by stripping header data and converting sequences to binary (e.g., formatdb of BLAST). Compression methods may also be useful here. In addition, the conversion process can be used as a way to gather information about the data so the software can be optimized automatically.

All methods that I am aware of build a one-dimensional array and scan *k*-mer matches. BLAST follows this up with a DPM, but most have found the results are problematic and too slow, so most eliminate the DPM. Choosing whether to build the array on the subject and scan the query or build on the query and scan the subject has profound implications for performance as well as for scoring. Searching with 25 nt words is important for a variety of reasons. This is not possible for SSAHA-based methods, but a method has been proposed to achieve 24 nt words by concatenating a pair of 12 nt hits [13] with very little performance cost over simple 12-mers (although much slower than a true 25-mer). SLIM Search is able to search with any word size directly, providing gains in flexibility, precision, and speed.

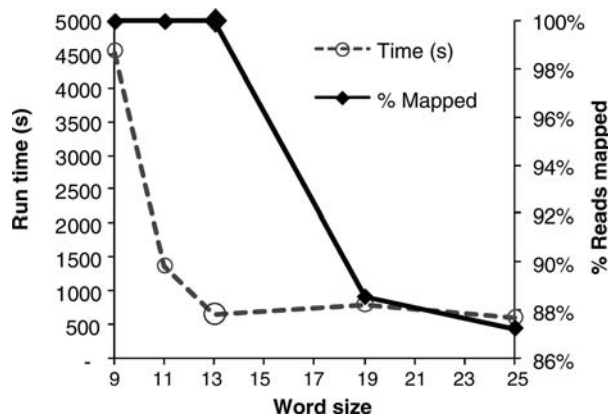


Figure 6.1 Optimization of word size for searching short NGS reads, demonstrating an optimum of short time and high hit rate at word = 13 for this data set.

Strategies that focus entirely on word matches, such as SLIM Search or BLAT, are becoming increasingly popular. However, it is important to understand how different parameters will affect results with NGS data. As word size increases, search speed increases at the expense of sensitivity, and the optimum will vary with read length and error rates. The effect of word size can be seen in Figure 6.1 where SLIM Search exhibits a clear optimum at word = 13 for mapping 100% of 1.5 million 454 reads to the yeast genome in about 11 min on a dual CPU Linux PC with a 2 GB RAM. At these conditions, about five positions are returned for each read. This can be reduced by introducing a threshold for a minimum number of Word Hits for each hit to be reported (MinHits). The effect of introducing a quality threshold can be seen in Figure 6.2 where SLIM Search shows an optimum of MinHits = 4, which reduces

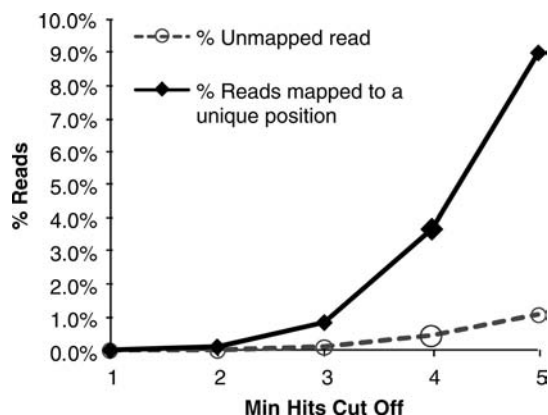


Figure 6.2 Optimization of MinHits filter for searching short NGS reads, demonstrating an optimum of increased unique mapping and minimal data loss at MinHits = 4 where 80% of spurious hits are filtered out.

hits to an average of about 2 per read and still maps virtually 100% of reads. This eliminates essentially all the millions of spurious hits, and you are left to resolve the true mapping with a Top-Hit filter from a few reasonable alternatives.

All current methods try to complete a search as a single operation; however, the speed of methods like SLIM Search makes a layered approach more attractive. It appears possible to improve performance by solving sections in layers, with improvements in both performance and precision. In addition, a layered approach makes it possible to drill down into data better such that every layer is dynamic and also quickly computed and displayed.

6.6

Subsequent Data Analysis

This discussion has dealt with the preliminary search, but a search is just the beginning of a project. There are hardware issues, and all the major hardware suppliers have offerings targeting NGS, with low-cost clusters looking to be the most popular. Subsequent analysis will require innovation in fundamental things such as file structures and database architectures. It is not clear whether the current RDBMS model can handle the magnitude and complexity of NGS data, or if a new paradigm is required. Subsequent analysis will require new tools both at the enterprise level and at the desktop level. NGS projects focus on relationships in large groups of data. The challenges of focusing key data on the researcher's desktop while maintaining important (undiscovered) relationships in the larger data set are not trivial. The key problems will revolve around the magnitude of the data: how to manage it so that the individual scientists can focus on what they need without losing key interactions and how to represent it so that complex data can be visualized without oversimplifying important details.

Most current NGS users are focused on metagenomics and epigenomics, but the obvious endgame for NGS is personalized medicine and the computational challenges for that are not trivial. Someone will have to build an international data repository with the entire genome and all relevant medical history for every person on the planet (or at least the First World), with access and security issues addressed. Every diagnostic lab will require software to deposit and update data, and every hospital will require software to integrate this information. Every medical practitioner will require a software interface to access this data, alongside a repository of clinical information to interpret it. In addition, many people will demand access to their own medical record, creating new computational challenges to provide medical information to individuals in a responsible and helpful manner so people can take an active role in their own wellness.

Traditional computational methods have not been working for NGS data, but there is a lot of promising research. New search technologies such as BLAT and SLIM Search are reducing search times from decades to days, or even minutes with SLIM Search. Computational improvements of this magnitude are helping to make personalized medicine a reality and realize the genomics dream for the benefit of humanity.

References

- 1 Quinlan, A.R., Stewart, D.A., Stromberg, M.P. and Marth, G.T. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods*, **2**, 179–181.
- 2 Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P. and Batzoglou, S. (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE*, **5**, e484.
- 3 Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Research*, **11**, 1697–1706.
- 4 Chaisson, M.J. and Pevzner, P.A. (2007) Short read fragment assembly of bacterial genomes. *Genome Research*, **2**, 324–330.
- 5 Warren, R.L., Sutton, G.G., Jones, S.J. and Holt, R.A. (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **4**, 500–501.
- 6 Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L. and Jones, C.D. (2007) *Bioinformatics*, **23**, 2942–2944.
- 7 Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.
- 8 Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.
- 9 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- 10 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.
- 11 Ning, Z., Cox, A.J. and Mullikin, J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Research*, **11**, 1725–1729.
- 12 Kent, W.J. (2002) BLAT: the BLAST-like alignment tool. *Genome Research*, **12**, 656–664.
- 13 Kalafus, K.J., Jackson, A.R. and Milosavljevic, A. (2004) Pash: efficient genome-scale sequence anchoring by Positional Hashing. *Genome Research*, **14**, 672–678.
- 14 Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, **30**, 2478–2483.
- 15 Giladi, E., Walker, M.G., Wang, J.Z. and Volkmoth, W. (2002) SST: an algorithm for finding near-exact sequence matches in time proportional to the logarithm of the database size. *Bioinformatics*, **18**, 873–877.
- 16 Ma, B., Tromp, J. and Li, M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
- 17 Bray, N., Dubchak, I. and Pachter, L. (2003) AVID: a global alignment program. *Genome Research*, **13**, 97–102.
- 18 Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A. and Batzoglou, S. (2003) NISC Comparative Sequencing Program. LAGAN and multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, **13**, 721–731.
- 19 Wu, T.D. and Watanabe C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
- 20 Kindlund, E., Tammi, M.T., Arner, E., Nilsson, D. and Andersson, B. (2007) GRAT: genome-scale rapid alignment tool. *Computer Methods and Programs in Biomedicine*, **1**, 87–92.

- 21 Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203–214.
- 22 Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human–mouse alignments with BLASTZ. *Genome Research*, **13**, 103–107.
- 23 SLIM Search Inc., 25108-B Marguerite Pkwy, No. 506, Mission Viejo, CA 92692, USA, unpublished. <http://www.slimsearch.com>.
- 24 Brenner, S.E., Chothia, C. and Hubbard, T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 6073–6078.
- 25 Korf, I., Yandell, M. and Bedell, J. (2003) *BLAST*, O'Reilly & Associates, Inc.
- 26 Blaisdell, B.D. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 5155–5159.
- 27 Lippert, R.A., Huang, H. and Waterman, M.S. (2002) Distributional regimes for the number of k -word matches between two random sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 13980–13989.
- 28 Lippert, R.A., Zao, X., Florea, L., Mobarry, C. and Istrail, S. (2004) Finding anchors for genomic sequence comparison. *Proceedings of the 8th Annual International Conference on Research in Computational Biology (RECOMB '04)*, ACM Press, 233–241.

7

DNASTAR's Next-Generation Software*Tim Durfee and Thomas E. Schwei*

7.1

Personalized Genomics and Personalized Medicine

The pursuit of personalized medicine and other newly emerging applications of next-generation DNA sequencing technologies require that the genome, or at least critical regions of it, be characterized at the nucleotide level. The need for this level of knowledge and understanding stems from the fact that even small-scale changes, including single-nucleotide polymorphisms (SNPs), can affect an encoded protein's function or the activity of a regulatory site. Such perturbations can have important health implications, including predisposing an individual to certain diseases, affecting his response to specific drug treatments, and altering his sensitivity to drug toxicities [1].

Molecular biology studies over the past 40 years have identified key causes of numerous diseases and served as the basis for understanding and/or developing efficacious drug treatments. Coupled with the ability to specifically target and sequence critical regions of an individual's genome, the stage is set to begin tailoring preventive and postclinical disease treatments to an individual's genotype. Indeed, till the time this chapter was written, several companies (23andMe, deCODEme, and Navigenics, among others) have begun offering personalized genomic services. These companies currently analyze approximately 1 million SNPs from an individual's genome (currently priced at about US\$ 1000). Nearly complete personal genome sequencing and analysis has also become available from Knome for US\$ 350 000.

As whole human genome sequencing becomes more affordable, dissecting the genetic basis of virtually any disease and/or drug efficacy/toxicity based on thousands of individual genomes has the potential to completely change the way medicine is practiced.

7.2

Next-Generation DNA Sequencing as the Means to Personalized Genomics

Driving the surge in momentum toward this new medical era are the next-generation DNA sequencing technologies described in the preceding chapters. Roche/454 Life

Sciences (454), Illumina, Applied Biosystems (AB), and polony sequencing each provide a cost-effective, massively parallel means of generating a vast amount of sequence data compared to traditional dideoxy (“Sanger”) sequencing. For example, the first human genome was determined at a cost of nearly US\$ 3 billion over the course of 10 years using strictly Sanger-based methods. The second “complete” genome, Dr James Watson’s, was recently completed for US\$ 2 million using Roche/454 technology (<http://www.technologyreview.com/Biotech/18809/>). As innovation and competition continue to improve sequence quality and yield, the goal of the US\$ 1000 personal genome announced by the National Institutes of Health (NIH) (<http://nihroadmap.nih.gov/>) will come closer to being a reality. That milestone, however, may well require one or more of the emerging technologies discussed in subsequent chapters.

In addition to personalized genomics, next-generation DNA sequencing technologies are also revolutionizing many other areas of biology. Microbial genomes can be completely sequenced with data from a single machine run of any of the current next-generation technologies, making epidemiological and evolutionary studies within reach of most reasonably funded laboratories. Community or metagenomic projects are now far more feasible than with Sanger-only technology. For example, a project to sequence the human microbiome is underway that has direct implications for personalized medicine as well [2]. Next-generation technologies are also augmenting traditional microarray-based applications, such as gene expression profiling, gene discovery, and gene regulation.

7.3

Strengths of Various Platforms

Each of the currently marketed next-generation sequencing platforms has unique advantages and disadvantages in various applications. The technologies can generally be divided into two groups: (1) the “long-read” (200–300 bases per read) Roche/454 technology that produces approximately 400 000 reads per machine run and (2) the “short-read” (25–50 bases per read) group of Illumina, AB, and polony sequencing that generate millions to tens of millions of reads per run. The Roche/454 “long-read” technology is best suited for *de novo* genome sequencing projects, whereas the “short-read” technologies are generally better suited for templated sequencing or resequencing projects, supporting SNP detection and discovery, digital gene expression, and similar applications.

7.4

The Computational Challenge

Unlike the uniformity and comparatively small data sets of the Sanger sequencing era, the diversity of next-generation technologies poses unique challenges to assembling and analyzing experimental results. Next-generation instrument

manufacturers provide sequence assembly software with their instrument that is tailored to the unique properties of their technology. However, this imposes unnecessary limitations on research in several ways. Mainly, combining different technologies in a single project is often desirable so that the strengths of one technology can compensate for the weaknesses of another. As mentioned above, investigators may use different technologies or combinations depending on the application, necessitating a flexible software solution. Postassembly, advanced data analysis tools that are integrated with the assembly software significantly streamline the data mining job. Finally, the cost-effectiveness of the next-generation technologies means that a far broader range of researchers can pursue genomic-scale approaches. To empower this trend, software should be easy to use and run on affordable computers. All these considerations point to a clear need for a third-party desktop software solution that supports all of the next-generation technologies and instrument providers.

7.5

DNASTAR's Next-Generation Software Solution

The desktop sequence assembly and analysis challenge has been an ongoing area of focus at DNASTAR since Lasergene, including the SeqMan assembly engine, was used to construct the *E. coli* genome in 1997 [3]. Since that initial success, SeqMan has evolved along two complementary tracks. SeqMan Pro remains effective as an assembler for smaller scale Sanger and Roche/454 sequencing projects and has developed into a comprehensive module for sequence assembly visualization, editing, and analysis within DNASTAR's Lasergene software suite. SeqManNGen Assembler (SNG) is DNASTAR's recently developed assembly engine that serves all next-generation technologies and effectively meets the increased needs for performance, capacity, and flexibility that these technologies demand, while still running on a desktop computer.

By adjusting assembly parameters to meet the needs of long- and short-read data sets, SNG can function with any individual platform data set or with a combination of data from multiple sequencing platforms. Moreover, next-generation data sets can be compared with any reference genome(s) allowing a straightforward extension of the assembler to resequencing projects. Running on a desktop computer, SNG has been used for the rapid, accurate assembly of numerous microbial genomes in both *de novo* and resequencing configurations. These projects have been completed with different next-generation sequencing platforms individually or in combination with each other and, at times, with data generated using Sanger sequencing (Table 7.1).

Following initial SNG assembly, projects are opened in SeqMan Pro, where contigs can be ordered into suprascaffolds using available mate pair data and with a genome aligner using an available reference sequence(s), if appropriate for the given project. At each stage of review and finishing, SeqMan Pro has several critical analysis tools including SNP reports in several different formats (Figure 7.1).

In areas of deep coverage, SNPs can be filtered in a variety of ways to reduce the volume of data viewed to a more manageable data set. Feature annotations can also be

Table 7.1 Examples of SMGA microbial genome assemblies.

Organism	Genome size (bp)	Platform(s)	Total number of reads	De novo/ resequencing	Data provider
<i>Escherichia coli</i> – MG1655	4 639 675	454	484 552	Both	454
<i>E. coli</i> – DH10B	4 686 137	Sanger	183 544	De novo	BCM-HGSC ^a
<i>Francisella tularensis</i>	1 895 727	Sanger	66 668	Both	BCM-HGSC
<i>Staphylococcus aureus</i>	2 872 915	454	317 789	Both	BCM-HGSC
		Sanger + 454	384 457	Both	BCM-HGSC
		Sanger	38 830	Both	BCM-HGSC
		454	429 553	Both	BCM-HGSC
		Sanger + 454	467 568	Both	BCM-HGSC
		Sanger + Illumina	2 860 448	Both	BCM-HGSC
		Sanger + 454 + Illumina	3 290 816	Resequencing	BCM-HGSC
<i>E. coli</i> ^b	4 686 137	454	484 552	Resequencing	454
<i>Saccharomyces cerevisiae</i>	12 162 996	Illumina	4 491 227	Resequencing	Guttman ^c

^aGenerously provided by George Weinstock, Baylor College of Medicine Human Genome Sequencing Center.

^bMG1655 data set assembled against the DH10B reference genome.

^cGenerously provided by David Guttman, University of Toronto.

added to a contig from the corresponding region of a reference sequence, from a BLAST search or directly in SeqMan Pro by the user. Data can easily be exported from SeqMan Pro to other tools, as well, if further processing or analysis is desirable. Taken together, these products provide the basic analysis links between next-generation DNA sequence assemblies and personal genomics on users' desktop computers.

Of course, it is one thing to provide desktop tools to handle microbial genomes, but it is quite another task to provide effective desktop tools for dealing with one or more human genome size data sets. DNASTAR is in the process of addressing this issue by incorporating additional algorithms into SNG that will ensure that the program is fully scalable to any size project, including those for personalized human genomics and medicine.

Figure 7.1 Example of interactive assembly and analysis windows within SeqMan Pro. Following assembly, an SNP report (bottom window) can be generated and entries filtered and sorted as defined by the user. Entries contain basic information about an SNP including the SNP frequency for a specified base position and whether the SNP affects the amino acid sequence of an annotated gene. Selecting an SNP in the report window highlights that column in the sequence alignment window (top) allowing the user to evaluate the change. Additional information about a feature can be viewed by using the mouse to hover over the feature's display.

7.6

Conclusions

Next-generation DNA sequencing technologies are providing the data generation capabilities for a new era of genomics and personalized medicine. Just as these technologies have transformed the sequencing landscape, DNASTAR's next-generation software is designed to transform the users' ability to analyze their genomic data and make personalized medicine and other critical next-generation applications a reality for life scientists.

Acknowledgments

We are extremely grateful to Schuyler Baldwin, Dr Richard Nelson, and Dr Carolyn Cassel, the primary developers of SeqMan Pro and SeqManNGen, as well as Daniel Nash and James Fritz. We also thank Dr Frederick Blattner and John Schroeder, the cofounders of DNASTAR, for their insights and stimulating discussions. Portions of this work were supported by NIH grants R43/4 HG01893 and R43 GM082117-01 to DNASTAR.

References

- 1 Roden, D.M., Altman, R.B., Benowitz, N.L., Flockhart, D.A., Giacomini, K.M., Johnson, J.A., Krauss, R.M., McLeod, H.L., Ratain, M.J., Relling, M.V., Ring, H.Z., Shuldiner, A.R., Weinshilboum, R.M. and Weiss, S.T. (2006) *Annals of Internal Medicine*, **145**, 749–757.
- 2 Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R. and Gordon, J.I. (2007) *Nature*, **449**, 804–810.
- 3 Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) *Science*, **277**, 1453–1474.

Part Four

Emerging Sequencing Technologies

8

Real-Time DNA Sequencing

Susan H. Hardin

8.1

Whole Genome Analysis

Despite recent advances, whole genome sequencing is still in its infancy. The process involves extensive reaction manipulation and computational analysis and remains time consuming and prohibitively costly for routine use. These constraints prevent widespread research into the genomes of many organisms and greatly limit the market potential and commercialization of a range of products and services centered on ultrahigh-speed DNA sequencing technology.

To rapidly advance biomedical research, a sequencing technology is required that can deliver cutting-edge speed at a much lower cost without sacrificing the high accuracy of current methods. A successful technology will produce DNA sequence information at the single-molecule level and in massively parallel arrays. This technology will lead to unknown applications and exponential growth in the nascent whole genome analysis market.

8.2

Personalized Medicine and Pharmacogenomics

There is an enormous, untapped market in both the medical and pharmaceutical industries for assessing disease risk and tailoring drug therapies for each patient's unique genome sequence.

Personalized medicine is the use of a patient's genotype or gene expression and clinical data to select a medication, therapy, or preventive measure that is particularly suited to that patient. The benefits of this approach are its accuracy, efficacy, safety, and speed. Personalized medicine will transform the medical profession – making it proactive, rather than reactive. Predicting and monitoring an individual's health by considering his or her genetic information and intervening before the individual becomes symptomatic will improve the overall health of the individual.

Pharmacogenomics enables scientists to search for and identify biomarkers that give clues as to how an individual will respond to a particular drug, incorporating information about a person's genetic makeup to develop a patient-specific treatment that ensures maximum efficacy with minimal adverse effects. In addition, pharmacogenomics can improve the understanding of disease mechanisms and facilitate identification of drug targets. Furthermore, drugs can be brought to market that improve the health of a subset of the population *and* that benefit the pharmaceutical industry. Pharmacogenomics will allow physicians to distinguish individuals who would benefit from a particular drug from those for whom the drug would be toxic. A best selling, efficacious drug could be kept on the market even if it is harmful to a small percentage of the population *if the cause of the adverse reaction has an understood genetic basis and a genomic analysis of each user is performed before the drug is prescribed by the physician*. Similarly, a drug that improves the health of participants during an early stage of a clinical trial may fail at a later stage. Currently, this outcome results in an essentially total financial loss of the investment to develop the drug. However, this drug may not become a total loss if the genetic bases for the beneficial and harmful effects of the drug are understood.

Drug dosage and regimen may be similarly tailored to patients' genomic information. The blockbuster business model of drug development, where drugs are prescribed in doses that work well for most, is being displaced by the pharmacogenomics approach. Whole genome sequencing technologies will unlock the full potential of personalized medicine and pharmacogenomics.

8.3

Biodefense, Forensics, DNA Testing, and Basic Research

Additional examples of potential uses for sensitive, ultrahigh-throughput sequencing include real-time identification and characterization of pathogenic organisms or genetically engineered biological warfare agents for which no (sequence) information is available. Forensic identification, paternity testing, transcriptome characterization, basic research applications, and animal, plant, and microbe genome characterization of organisms are additional uses for low-cost, ultrahigh-throughput whole genome sequencing technologies.

8.4

Simple and Elegant: Real-Time DNA Sequencing

VisiGen Biotechnologies, Inc. is developing a breakthrough technology to sequence a human genome in less than a day for less than \$1000. The ability to achieve \$1000 human genome sequencing is directly related to the successful implementation of a single-molecule approach, and the ability to accomplish this feat in a day is directly related to the implementation of the approach on a massively parallel scale.

VisiGen is distinguished from other leading developers of next-generation sequencing technologies in that it exploits the natural process of DNA replication in a way that enhances accuracy and minimally impacts efficiency. VisiGen has engineered both polymerase, the enzyme that synthesizes DNA, and nucleotides, the building blocks of a DNA strand, to act as direct molecular sensors of DNA base identity in real time, effectively creating “nanosequencing machines” that are capable of determining the sequence of any DNA strand. The technology platform detects sequential interactions between a single polymerase and each nucleotide that the polymerase inserts into the elongating DNA strand. Importantly, before initiating sequencing activity, the nanosequencers are immobilized on a surface so that the activity of each can be monitored in parallel.

VisiGen’s scientists build nanosequencers by drawing on the disciplines of single-molecule detection, fluorescent molecule chemistry, computational biochemistry, and genetic engineering of biomolecules.

VisiGen’s strategy involves monitoring single-pair Förster resonance energy transfer (spFRET) between a donor fluorophore attached to or associated with a polymerase and a color-coded acceptor fluorophore attached to the (terminal) γ -phosphate of a dNTP during nucleotide incorporation and pyrophosphate release (Figure 8.1). The purpose of the donor is to excite each acceptor to produce a fluorescent signal for which the emission wavelength and intensity provide a unique signature of base identity.

Working at the single-molecule level makes both sequencing signal maximization and background noise minimization critical. Thus, the core of VisiGen’s technology exploits aspects of physics that enhance signal detection to enable real-time, single-molecule sequencing. First, VisiGen’s spFRET approach is directly influenced by distance – both for maximizing signal and minimizing background noise – because FRET efficiency is an inverse function of the sixth power of the distance between donor and acceptor fluorophores [1–7]. In particular, the molecules are engineered to maximally FRET when the acceptor-tagged dNTP docks within the polymerase’s active site, thereby maximizing the acceptor signal indicating base identity, minimizing acceptor emission until the acceptor fluorophore is sufficiently close to a donor fluorophore to accept energy, and producing an anticorrelated change in donor intensity – all of which improve incorporation event detection. Second, the sequencing platform incorporates total internal reflectance fluorescence (TIRF) to further increase the signal-to-noise ratio by minimizing the depth of light penetration. This approach is effective at minimizing background noise because most of the labeled dNTPs in the reaction solution are not within the TIRF excitation volume and are, therefore, not excited by the incident light.

Donor fluorescence is both informative and confirmatory because it is anticorrelated with acceptor fluorescence throughout the incorporation reaction. After an spFRET event, the donor’s emission returns to its original state and is ready to undergo a similar intensity oscillation cycle with the next acceptor-tagged nucleotide. In this way, the donor fluorophore acts as a punctuation mark between nucleotide incorporation events. The donor fluorophore’s return to its pre-spFRET intensity between incorporation events is especially important during the analysis of homopolymeric sequences.

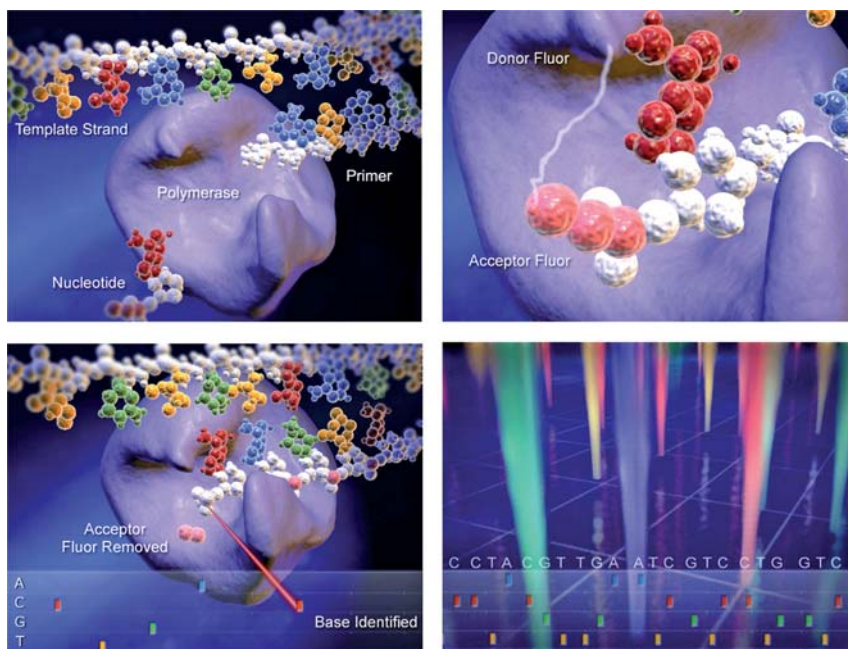


Figure 8.1 Real-time DNA sequencing. Top, left: Reaction components of VisiGen's sequencing system include modified polymerase and nucleotide, primer, and template. Top, right: Energy transfers from the donor fluorophore to the acceptor fluorophore on the gamma-dNTP, stimulating acceptor emission and sequence detection. Bottom, left: Fluorescently tagged PP_i leaves the nanosequencing machine, producing natural DNA. A noncyclic approach enables rapid detection of subsequent incorporation events. Bottom, right: Arrays of nanosequencing machines. The time-dependent fluorescence signals emitted from each asynchronous sequencing complex are monitored and analyzed to determine DNA sequence information. Massively parallel arrays enable ultrahigh throughput (1 million bases per second per machine).

During an extension reaction, when a nucleotide is incorporated into the growing DNA strand, energy transfers from the polymerase to the nucleotide via spFRET, thereby stimulating the emission of a base-specific incorporation signature that is directly detected in real time. During nucleotide insertion, the 3' end of the primer attacks the alpha-phosphate within the dNTP, cleaving the bond between the alpha- and beta-phosphates and also potentially changing the spectral properties of the fluorophore (which remains attached to the released pyrophosphate; PP_i). In addition, because the nucleotides are fluorescently modified at the γ -phosphate, VisiGen's approach produces a native DNA polymer, rather than a highly modified polymer that would negatively impact polymerase activity, and facilitates incorporation of sequential gamma-labeled nucleotides. Importantly, this approach actually enables data collection *during* the sequencing reaction, with the associated benefits of eliminating the need for subsequent reagent deliveries and reducing data acquisition time due to the coupling of the synthesis and acquisition phases of the reaction.

VisiGen's core technology associates the donor with an immobilized component of the nanosequencer to minimize background fluorescence. The technology associates the donor with an immobilized polymerase to obtain consistent signals, rather than at a specific site on the primer/template (which increases the distance between the donor and the acceptor with each nucleotide insertion and produces less consistent signals). A donor-tagged, immobilized polymerase maintains a constant distance between the donor and the acceptor during nucleotide incorporation, producing high FRET with consistent intensity signatures, and positions the nanomachine within the illuminated volume at a relatively constant and higher energy position near the surface. Together, these consistencies minimize data analysis complexity and facilitate longer sequence reads.

Existing technologies enable the creation of grids of 1000 of these nanosequencers in a single field of view, and each sequencing cassette contains over a billion nanosequencers. The massively parallel approach incorporated into the VisiGen sequencing system will enable the company to produce an instrument that is controlled by a single operator and collects sequence data at a rate of 1 Mb/s/machine, approximately 100 Gb of DNA sequence information per day.

Acknowledgments

Development of VisiGen's sequencing technology is funded by grants and contracts from DARPA and NIH and by private investments from SeqWright, Inc. and Applied Biosystems. Discussions with VisiGen development teams are gratefully acknowledged, as is assistance from Christopher Hebel.

References

- 1 Förster, T. (1948) Zwischenmolekulare energiewanderung und fluoreszenz. *Annalen der Physik*, **2**, 55–75.
- 2 Stryer, L. and Haugland, R.P. (1967) Energy transfer: a spectroscopic ruler. *Proceedings of the National Academy of Sciences of the United States of America*, **58** (2), 719–726.
- 3 Stryer, L. (1978) Fluorescence energy transfer as a spectroscopic ruler. *Annual Review of Biochemistry*, **47**, 819–846.
- 4 Dale, R.E., Eisinger, J. *et al.* (1979) The orientational freedom of molecular probes. The orientation factor in intramolecular energy transfer. *Biophysical Journal*, **26** (2), 161–193.
- 5 Clegg, R.M., Murchie, A.I. *et al.* (1993) Observing the helical geometry of double-stranded DNA in solution by fluorescence resonance energy transfer. *Proceedings of the National Academy of Sciences of the United States of America*, **90** (7), 2994–2998.
- 6 Selvin, P.R. (2000) The renaissance of fluorescence resonance energy transfer. *Nature Structural Biology*, **7** (9), 730–734.
- 7 Weiss, S. (2000) Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature Structural Biology*, **7** (9), 724–729.

9

Direct Sequencing by TEM of Z-Substituted DNA Molecules

William K. Thomas and William Glover

9.1

Introduction

Sequencing is the most powerful form of genetic analysis. Widespread knowledge of human genetic variation has the potential to unlock the mysteries of the most complex human diseases, especially the myriad forms of cancer, and to foster genotype-specific drug therapies significantly impacting the quality of life. The first draft genome of an individual human diploid genome has just been published [1] and has revealed both the potential knowledge to be gained and the significant challenges to be faced by current sequencing approaches. The most obvious remaining challenges are cost and data quality. Sequencing costs remain prohibitive for any dramatic expansion of individual human genome sequencing, and although most current methods reduce costs and increase throughput, they also create limitations in assembly and analysis that constrain the discovery of many aspects of genetic variation due to the short-read lengths. The approach described in this chapter, direct sequencing by transmission electron microscopes (TEMs), represents a novel method that could not only generate sequence data at a fraction of current costs but also significantly improve data quality, allowing a more complete understanding of genetic variation within and among individuals.

ZS Genetics (ZSG) is developing a method to prepare DNA for sequencing by direct imaging with TEM. In this approach, single discrete DNA molecules can be imaged and individual base pairs identified, providing fast, automated results. With this approach, a sample is prepared once, a picture is taken, and the sequence is read directly from that picture. Compared to the fastest systems currently available that can assay 2–100 million bases in a day, sequencing by TEM has the theoretical potential to assay 40–100 million bases per hour. Moreover, this technology encompasses more than just increased sequence data production at reduced cost. The approach has the potential to produce very long, continuous sequence reads, at least several kilobase (kb) pairs long. The production of several kilobase long reads would fundamentally change the ability of researchers to more

fully assemble complex genomes, evaluate transcript splice variation, and accurately assess SNP associations.

9.2

Logic of Approach

DNA molecules are small. Nucleotide pairs are even smaller. All sequencing technologies strive to bridge the gulf between the macrorealm of human perception and subnanometer realm of nucleic acid information storage. TEMs work naturally at the scale of molecules and atoms. TEMs can create images with spatial resolution better than 1 Å, but have not been used to image individual base pairs because unlabeled nucleotides cannot be seen with previous TEM dyeing and labeling technologies. A TEM visualizes the charges on atomic nuclei, defined as the atomic number (Z). Atoms with low Z are transparent to TEM, and materials made of atoms with similar atomic numbers have no contrast. The average Z for natural DNA is around 5.5, and given the ladder-like structure of a double-stranded DNA (dsDNA) molecule more than half that volume is simply void, resulting in an effective average Z of about 2 with almost no difference from base to base. Consequently, natural DNA is essentially invisible to TEM analysis. While neither light microscopes nor electron microscopes can “see” natural DNA, even entry-level systems are routinely warranted by vendors to achieve point-to-point resolution of less than 1 nm. Midrange systems give results around 0.2 nm, and the most advanced systems can achieve resolution of less than 0.85 nm (Table 9.1). This spatial resolution is just over half the carbon–carbon bond length, less than half the hydrogen bond length, and one-quarter the typical distance between adjacent base pairs in a dsDNA molecule. Therefore, the problem faced by TEM imaging of DNA molecules is not resolution but contrast.

Our approach makes DNA visible by incorporating modified nucleotides into a DNA molecule that have one or more atoms substituted or an additional atom attached with a higher and unique Z number. In this way, if each of the four bases is both visible and distinguishable from one another, the sequence of a dsDNA molecule can be read from a TEM image.

A key to the success of this approach is the synthesis and incorporation of such Z -modified nucleotides (Z -dNTPs) into DNA molecules. As stated above, TEMs “see” nuclear charges (total number of protons, or Z) and the average effective Z of DNA is ~ 2 . However, using middleweight common modifications, for example, iodine

Table 9.1 Representative TEM performance specs.

System type	Vendor	Product	Point-to-point resolution
Basic	Hitachi	H-7650	0.36 nm
Mid-range	Zeiss	Libra 200MC	0.24 nm
Advanced	FEI	Titan 80–300	0.9 nm or better

Table 9.2 Potential modifications to create Z-dNTP library.

Base	Location on base	Substituent	Location on phosphate	Substituent
A	C7 (7-deaza-7-X), C8	I, Br	α P	S, Se
G	C7 (7-deaza-7-X), C8	I, Br	α P	S, Se
T	C5, C6	I, Br, CF ₃ , CCl ₃	α P	S, Se
C	C5, C6	I, Br	α P	S, se

(Z = 53) or bromine (Z = 35), adequate contrast can be obtained. While most of the unlabeled DNA molecule remains transparent, the individual, labeled atoms are visible. A library of such Z-modified nucleotides (Z-dNTPs) can be accomplished by substitutions at nondisruptive locations on the nucleotides (Table 9.2), and several are commercially available.

To incorporate the Z-tag nucleotides, we are proposing two rounds of linear DNA replication (linear amplification), one for each strand. In each case, unique mixtures of Z-dNTPs can be incorporated in each round. In all cases, replication is accomplished with complete substitution of specific Z-dNTPs. In the proposed process, the first strand is replicated several times producing an excess of single-stranded DNA (ssDNA) with labels on one or two of the nucleotide types. This DNA is then used as the template for the final set of linear amplifications. This final round using the modified ssDNA as a template will create dsDNA with a second group of modified dNTPs on the final strand. By using several rounds of synthesis, the vast majority of molecules will be labeled on both strands. As shown in Figure 9.1, by incorporating three different Z-dNTPs that are both visible and distinguishable by TEM, it is possible to unambiguously read the DNA sequence.

There are multiple key steps in this process that must be demonstrated and ultimately optimized for this technology to realize its full potential. First is to select a set of Z-dNTPs that maximize TEM-based visualization (unique contrast). These must also be efficiently incorporated by polymerases (see the next section). The final

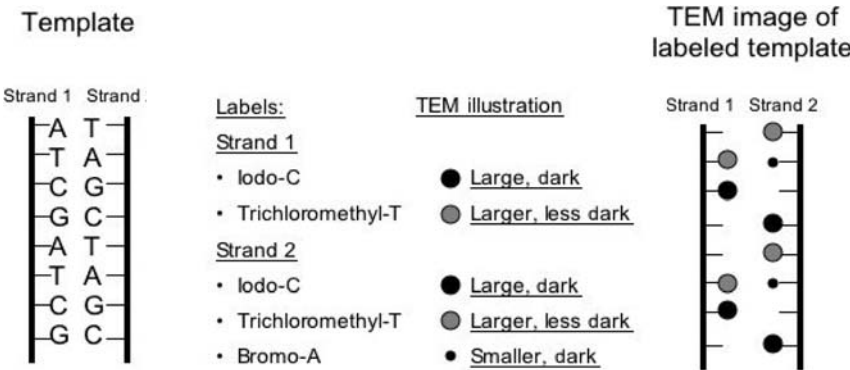


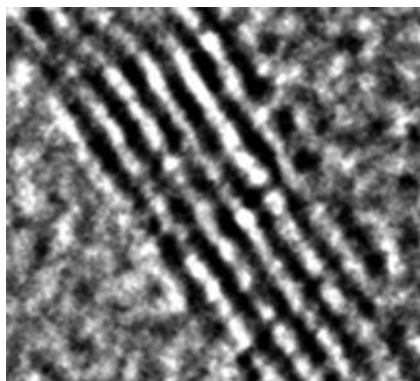
Figure 9.1 Labeling strategy with three Z-tagged nucleotides.

critical step in the process is the display of labeled dsDNA on a TEM substrate that maximizes the visual analysis.

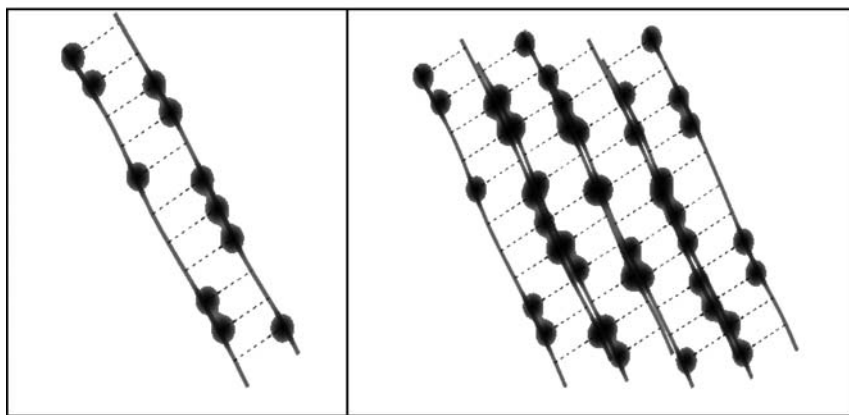
9.3

Identification of Optimal Modified Nucleotides for TEM Visual Resolution of DNA Sequences Independent of Polymerization

Preliminary studies have shown that it is possible to efficiently incorporate several commercially available dNTPs labeled with heavy atom labels (Z labels) even by enzymatic polymerization with standard DNA polymerases (Figure 9.2). The initial



(a)



(b)

Figure 9.2 (a) TEM images of stretched DNA with 5-iodo C and 5-iodo U nucleotides and (b) component molecules based on image analysis of aligned dsDNA molecules in (a). The image in (b) on the right shows the four dsDNA molecules with overlapping bases on adjacent molecules, and one of the molecules is shown on the left.

optimization steps focus on the empirical analysis of individual nucleotides and the ability to distinguish each by TEM to define a set of reagents that can be further optimized for maximal sequence lengths of DNA sequencing reads. At present, there is much interest in the use of modified nucleotides in DNA molecules, largely driven by the need to generate functional nucleic acids or “DNAzymes” [2, 3] and develop new methods for genotyping [4]. Studies on the incorporation of modified molecules suggest that a wide range of modified nucleotides can be incorporated into DNA by commercially available polymerases. Others clearly demonstrate that the efficiency of incorporation of specific modified nucleotides depends upon several variables including the enzyme used and the sequence of the template [5–8].

Table 9.2 illustrates the potential modifications that ZSG is interested in pursuing including the location of the substitution and the molecule. More than one substitution can be made per nucleotide, resulting in several possible permutations as Z-labeled nucleotides (Z-dNTPs).

9.4

TEM Substrates and Visualization

The final critical step in sequencing with the Z-tagged DNA is the effective localization of the labeled DNA molecules on an appropriate TEM substrate. Ideally, the molecules can be aligned via fluid flow along the imaging window. Alignment techniques such as molecular combing [9–11] and microfluidic alignment [12] will be used to stretch and align the molecules for viewing. Under specific conditions, these processes stretch the DNA molecule, changing it from a helix to a ladder conformation in at least local areas, as shown in our preliminary work (Figure 9.2).

In the images shown in Figure 9.2, we establish the feasibility of using Z-tag labels to visualize nucleotides with atomic resolution, using a TEM instrument and 5-iodo-dUTP- and 5-iodo-dCTP-substituted DNA molecules generated by primer extension. Figure 9.2a shows a region of an electron micrograph with five DNA molecules that have been completely substituted for two nucleotides. Based on the spacing of these images, the series of dots in ordered rows are the iodine atoms within the DNA molecules on 5-iodo-dUTP- and 5-iodo-dCTP. The white areas between the rows of dots are interstrand spaces. In this section, the molecules appear to have unwound from the standard double helix into a ladder conformation ideal for visualization. The iodine atoms, covalently attached to the 5-carbons, are very near to the edge of the molecules. Consequently, with the labels from one molecule very close to the labels of the next, they tend to blur together. However, analysis with fast Fourier transformation indicates that the primary repeating pattern is precisely 3.5 Å in pitch, slightly more than expected for A-form DNA. Note that in some of the interstrand space, faint lines can be seen that cross the white space at right angles. We believe these are those nitrogenous bases that happened to stabilize at right angles to the imaging beam, allowing multiple atoms in the planar rings to add together for contrast purposes. On the basis of this analysis, we can interpret the image to represent a set of parallel, double-stranded DNA molecules. In some cases, adjacent nucleotides have iodine-substituted atoms (Figure 9.2b).

Separation of one of the predicted dsDNA molecules suggests that the orientation of U and C residues is the same as predicted in dsDNA.

9.5

Incorporation of Z-Tagged Nucleotides by Polymerases

Toward the development of specific reagents for sequencing by TEM visualization of Z-modified DNA molecules, we have screened several Z-modified nucleotides for robust incorporation by commercially available DNA polymerases (and reverse transcriptase). In all, we have tested 14 unique modified nucleotides individually (Figure 9.3a–c and Table 9.3). The assay focuses on screening for efficient incorporation of modified nucleotides to support the PCR amplification process using primers that specify a 520 bp product. However, a standard set of template lengths (520 bp, 1011 bp, 2 kb, 3 kb, 5 kb, and 23 kb) was used in each assay; primers are given in Table 9.3. In many cases, much larger products were amplified using the primer pairs that amplify larger fragments. However, because the ultimate method of

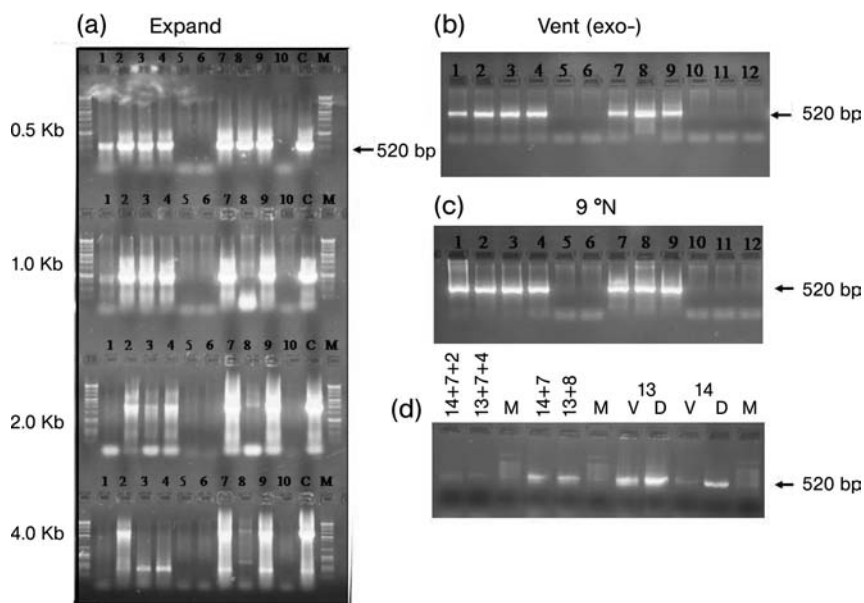


Figure 9.3 Gel electrophoresis of PCR amplifications using Z-modified nucleotides. (a) Assays using Roche Expand and 10 Z-modified nucleotides (Table 9.1). PCR products are shown for four different primer pairs using lambda DNA as template. (b and c) Twelve amplifications with unique Z-modified nucleotides and the 520 bp primer pairs.

(d) Single substituted amplifications of the Z-modified nucleotides 13 and 14 using either DyNAzyme (D) or Vent (V). Left side of the gel shows combinations of nucleotides used in successful amplifications of 520 bp product. In all gels, the Z-modified nucleotides are numbered as in Table 9.3.

Table 9.3 Z-modified nucleotides tested for incorporation in PCR assay.

No.	Z-nucleotide	Amplification	No.	Z-nucleotide	Amplification
1	dTTP-alphaS	Yes	8	5-Br-dCTP	Yes
2	dGTP-alphaS	Yes	9	5-I-dUTP	Yes
3	dATP-alphaS	Yes	10	5-I-dCTP	No
4	dCTP-alphaS	Yes	11	2'-I-ATP	No
5	8-oxo-dGTP	No	12	2'-Br-ATP	No
6	8-Br-dATP	No	13	7-Deaza-7-iodo-dGTP	Yes
7	5-Br-dUTP	Yes	14	7-Deaza-7-iodo-dATP	Yes

incorporating modified nucleotides to support TEM-based sequencing is not well modeled by a PCR reaction, no attempt was made to optimize the PCR reactions to achieve maximum product size. This assay is simply to screen modified nucleotides for efficient incorporation.

Table 9.3 lists the results for the first 14 nucleotides. In 9 of the 14 cases, an efficient amplification was observed. For each modified nucleotide, five different DNA polymerases were used in standard PCR reactions. These enzymes were Taq polymerase, DyNAzyme, Roche Expand, 9° North (exo-), and Vent (exo-) (see Table 9.2 for details). Some variability in the incorporation was observed among enzymes, and Taq DNA polymerase alone was the least efficient for the incorporation of modified nucleotides.

We consider this assay to be extremely stringent. Successful specific amplification by PCR requires much more robustness than we will use for commercial practice. Importantly, successive cycles of PCR require that the modification be tolerated both for incorporation during strand synthesis and in the template strand in successive rounds. As discussed above, in practice, labeling will have a first set of reactions to incorporate only those labels that are well tolerated in the template strand; a final round will incorporate those that are tolerated in the synthetic strand but not in the template strand. In this way, those modified dNTPs that are inhibitory when read as a template will not be used in the template strand.

9.6

Current and New Sequencing Technology

There are over 6 billion base pairs in a diploid human genome. To identify each nucleotide uniquely, *something* has to be done 6 billion times. The majority of recent sequencing innovators use some form of biochemistry, then pause to collect data to identify each of those base pairs. This creates 6 billion cycles of biochemical reactions and data collection. And this is before the redundancy that is generally required for accuracy. The reactions for labeling DNA with Z-modified nucleotides are done only once. We do our chemistry once and therefore do not require a pause between base pair identifications. We will identify tens of thousands of base pairs in each image, pausing only to take another image. Moreover, image processing software can

identify the base pairs in real time. By the time the imaging is finished and the sample removed from the machine, the entire data set will have been processed. This is aided by the ability to label and image molecules that are tens of thousands of base pairs long, which also reduces the need for very high levels of coverage required in some other approaches generating extremely short reads.

Capillary electrophoresis is the incumbent technology. This method requires millions of repetitive chemistry steps along with processing hundreds of thousands of samples, limiting the cost reduction to that which automation and miniaturization can achieve. A strength of this technology is that it captures contiguous sequence reads of 800–1000 bases, which makes assembling a whole genome much easier than the much shorter reads of most innovative technologies. However, improvements are not viewed as having the practical potential to achieve industry benchmarks of \$100 000 and \$1000.

Sequencing-by-synthesis approaches have the common characteristic of reading bases as they are added to DNA strands by using fluorescent tags excited by lasers or light emitting diodes. The general advantage of this approach is massively parallel processing. This helps them cope with the need to perform billions of chemical reactions per genome. Small sample size also reduces the use of expensive reagents and supplies. The main disadvantage is short-read lengths, from as few as 35 to as many as a few hundred base pairs. ZSG considers sequencing-by-synthesis technologies to be extremely fast, but not much cheaper than traditional methods. Companies involved in sequencing by synthesis will be very successful in applications that aim at a very small fraction (less than 1%) of a human genome (e.g., viruses and bacteria) and where speed is far more important than cost. 454 Life Sciences Corporation, a subsidiary of Roche Appl Science, was first to market; ZSG considers them the best of the category.

Single-molecule sequencing technologies are often similar to sequencing by synthesis but do not rely as much on amplification. These technologies have the advantage of massively parallel processing over capillary electrophoresis but have even more substantial problems with short-read lengths. So far, they cannot read more than 50 bases at a time. This creates an enormous problem with reconstructing the original sequence from these short sequences. Paired end reads will help considerably but at a substantial penalty in overall costs and total process speed.

Nanopore sequencing technologies hope to isolate individual DNA strands and pass them through a pore that reads the bases electrically or optically. This interesting approach is still in infancy at university labs, with a dozen or more viable teams. Many scientists view nanopore sequencing as having tremendous potential, especially for parallel processing. Most also believe this technology to be 10 years from reaching the market. Breakthroughs are especially needed for the challenges of physically processing DNA strands through 1–2 nm structures and reading mechanisms.

Unlike the current and innovative sequencing approaches discussed above, ZSG's TEM sequencing technology will have extremely low operating costs (reagents and other consumables are negligible), long-read lengths (>15 kb), and high throughput (up to 100 million bases per hour).

9.7

Accuracy

A critical aspect of DNA sequencing, especially in the context of linking DNA polymorphisms and human health, is accuracy. There are two distinct and inherent sources of error in sequencing a genome, accuracy of base calls including polymorphism and accuracy of the assembly. In a real-world application to a single human diploid genome, the accuracy must allow the identification of heterozygotes and variation in the common human repeat classes. The proposed direct sequencing by TEM technology has two advantages with respect to accuracy. First, with respect to base calling accuracy, the methodology proposed here allows the analysis of multiple independent representations of each section of a DNA molecule. As with current approaches, the most powerful improvement in DNA sequence quality with respect to base calling comes from the fact that each nucleotide is called multiple times and the very low error rates are multiplicative. Second, the extremely long reads produced by this methodology significantly reduce the extent of paralogy. Paralogy, or recently duplicated segments of the genome, represents a major problem in resequencing. These recently duplicated segments comprise a significant fraction of the human genome, have played a pivotal role in the structural evolution of human chromosomes, and contribute disproportionately to a large number of genetic diseases [13, 14]. In fact, a significant percentage (>5%) of the human genome contains segments greater than 5 kb and 90% identity [15].

Traditional sequencing methods (<1 kb) [16] and the shorter reads (50–250 bp) produced by recently developed sequencing strategies [17, 18] are problematic with respect to these segmental duplications because they are hard to uniquely assign to specific loci in a reference sequence and virtually impossible to accurately assemble. By contrast, TEM-based reads of >10 kb will have a much higher probability of uniquely matching sequence in the human genome that will make it easier to assay variation in paralogous regions that contribute to human disease.

The biggest question mark for TEM sequencing accuracy is the use of halogenated nucleotides, especially the ones that are explicitly known to be mutagenic in normal PCR reactions. ZSG plans to overcome this problem by using only two rounds of amplification, and with the most mutagenic nucleotides only in the final round. In this way, the mutagenic nucleotides will not ever be used as a template for further synthesis.

9.8

Advantages of ZSG's Proposed DNA Sequencing Technology

1. The ability to rapidly and directly sequence entire genomes without the production of clone libraries.
2. The requirement of only a few cycles of PCR (versus 30 +), eliminating most of the chemistry and most of the error that PCR induces.

3. Increased accuracy by taking advantage of the inherent redundancy of double-stranded DNA molecules and the ability to sequence both strands at one time.
4. Low limits of detection; it is expected that less than 50 copies of each DNA strand should be required.
5. The ability to sequence individual molecules and assemble long continuous haplotypes, revealing heterozygosity, the key to real genome-wide personal sequencing.
6. Read lengths of 10 000–20 000 bp reduce or eliminate paralogy, facilitate data assembly, and increase the probability of uniquely matching imaged sequences with known databases.
7. High throughput with potentially hundreds of millions of bases per hour, or billions per day.
8. Fast turnaround time. It is expected that the entire human genome can be sequenced in a day or less including the sample preparation.
9. Instrument development for commercialization requires modification to existing TEM technology to reduce cost and increase throughput, not developing the instrument.
10. Minimal software development required.

9.9

Advantages of Significantly Longer Read Lengths

Read lengths of several kilobases offer distinct advantages in multiple aspects of genome analysis including the assembly of *de novo* genomes, sequence-based transcriptome analysis, and genome-wide analysis of variation.

9.9.1

***De novo* Genome Sequencing**

Current methods produce sequence reads ranging from 36 to 800 bp. These reads comprise the first step in whole genome assembly through unique matching of overlapping sequences resulting in a tiling path of reads to produce contigs. In theory, even large genomes with high sequence complexity (nonrepetitive) can be assembled *de novo* from relatively short reads (<50 bp); however, the contigs resulting from real genome sequence assemblies are directly affected by both the read length and complexity of the genome [19, 20]. The efficiency of the assembly ultimately boils down to two key issues that depend on read length. First, the matches between reads must be unique, and the probability of a sequence read being unique depends on length even in a nonrepetitive genome. As discussed below, this unique matching issue is also critical to the correct assembly of individual haploid genomes from a set of sequence reads in a heterozygous diploid individual. The second major challenge is directly related to the inherent repetitive nature of the genome being sequenced. All genomes, especially large, complex genomes such as a human genome, consist of many segments that are indistinguishable at the scale of traditional sequences. For

Table 9.4 Repeat elements in the human genome [1].

Repeat class	Number	Average length
SINEs	1 738 571	227
LINEs	957 647	632
LTR elements	474 016	517
DNA elements	307 288	276
Unclassified	5217	434
Small RNA	11 049	128
Satellites	93 568	1106
Simple repeats	447 165	68
Low complexity	380 093	46

example, highly repetitive sequences hundreds of nucleotides long comprise almost 50% of the human genome (Table 9.4). All of these major classes of repetitive DNA severely restrict contig formation by reads in the traditional (800 bp) and smaller size ranges. Furthermore, in addition to the classical highly repetitive elements in complex genomes, recent gene duplication and structural variation is much more widespread than expected [21–24]. These recent segmental copy number changes will confound assembly. Consequently, while it is possible to assemble contigs for the nonrepetitive regions of the genome, these widespread, recently duplicated regions show copy number variation that exceeds the amount of genetic variation associated with SNPs and span 12% of the genome including hundreds of genes [25].

To overcome these problems, traditional clone-based genome sequencing has incorporated the use of paired end reads and BAC fingerprinting to span repetitive regions and uniquely resolve the order and orientation of sequences in complex genomes. Similar approaches to overcome the restriction to assembly but without the generation of traditional clone libraries have also recently been developed for 454-based analysis [24, 26] and Helicos BioSciences (<http://www.helicosbio.com/>). But these approaches are currently limited to paired end reads that are extremely short in sequences and span only a restricted distance. Although this general strategy significantly improves the ability to assemble scaffolds of contigs in the correct order and orientation and assay structural variation, it remains limited by the uniqueness of the individual reads, a process that strictly depends on sequence length.

9.9.2

Transcriptome Analysis

An important traditional application of DNA sequencing is the description and analysis of the transcriptome. DNA sequence analysis of a transcriptome helps define gene structure by providing unambiguous evidence of splice junctions [27] and can form the basis of transcriptome profiling. Since the advent of serial analysis of gene expression (SAGE [28]), direct sequence analysis has also contributed to the knowledge of gene expression patterns. Direct sequence analysis of transcript representation has largely given way to hybridization-based methods [29, 30].

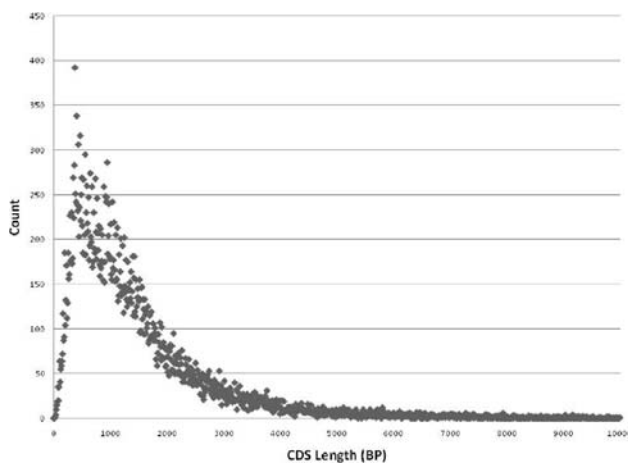


Figure 9.4 Size distribution of human coding sequences.

New massively parallel sequencing technologies, because of their large-scale and theoretical representation of single molecules, have the potential to return the analysis of gene expression to a sequence-based approach [31–34].

The analysis of a transcriptome also depends upon read length. The generation of cDNA sequences from ESTs either by traditional sequencing of cDNA libraries or by more recent approaches using innovative technologies depends upon successful assembly/clustering of reads and is limited especially by recent gene duplicates (paralogs) as discussed above. More importantly, however, because such a large proportion of mammalian genes are alternately spliced, knowledge of the complete mature mRNA sequence and any variants is critical to the characterization of the transcriptome and to the evaluation of expression dynamics. Short reads and indirect hybridization-based analysis do not provide the direct knowledge of transcripts spanning the entire coding sequences.

An analysis of human transcript lengths based on the size of the coding regions (CDSs; Figure 9.4) shows that the vast majority of transcripts exceed 1 kb, a severe challenge to new short-read technologies. However, 95% of transcript CDSs are less than 5 kb suggesting that even read lengths of 5 kb will allow a detailed analysis of the vast majority of the protein coding portions of human transcripts. The direct visualization of DNA sequences by TEM has the potential to allow sequencing, for the first time, of complete coding portions of transcripts and any alternately spliced forms.

9.9.3

Haplotype Analysis

A third aspect of the human genome sequence analysis that can benefit greatly from longer read lengths is the analysis of human genetic variation. The basic goal of human genome sequencing is to associate human phenotypic variation (disease) with genetic variation. Because humans are diploid, we ultimately must describe human genetic variation in the context of each of the two haploid genomes and derive

individual human genome sequences from diploid mixtures. In the extreme sense, the sequences of the individual chromatids must be assembled out of the sequencing reads. In a recent analysis of an individual human diploid genome, Levy and colleagues [1] were able to generate large contigs representing individual haploid regions of the human genome using 7.5-fold coverage of traditional reads. In their novel haplotype assembly, they were able to assemble approximately half the human genome into haplotype-specific segments >200 kb. One key to the efficient assembly of haploid genomes or haplotypes is that reads span multiple variant nucleotide positions between the two alleles. The average distance between SNPs as given by the nucleotide diversity [35] within the human reference sequence [1] is 6.15×10^{-4} or approximately one heterozygous position every 6000 nucleotides, an estimation consistent with other studies [36]. With spacing of heterozygous loci on average thousands of nucleotides apart, read-based contig assembly will require much longer reads to significantly improve haplotype-specific assembly.

References

- 1 Levy, S. *et al.* (2007) The diploid genome sequence of an individual human. *PLOS Biology*, **5**, 2113–2144.
- 2 Breaker, R.R. (1997) DNA aptamers and DNA enzymes. *Current Opinion in Chemical Biology*, **1**, 26–31.
- 3 Ting, R.T. *et al.* (2004) Selection and characterization of DNazymes with synthetically appended functionalities: a case of a synthetic RNaseA mimic. *Pure and Applied Chemistry*, **76**, 1571–1577.
- 4 Wolfe, J.L. *et al.* (2002) A genotyping strategy based on incorporation and cleavage of chemically modified nucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11073–11078.
- 5 Giller, G. *et al.* (2003) Incorporation of reporter molecule-based nucleotides by DNA polymerases. I. Chemical synthesis of various reporter group-labeled 2'-deoxyribonucleoside-5'-triphosphates. *Nucleic Acids Research*, **31**, 2630–2635.
- 6 Tasara, T. *et al.* (2003) Incorporation of reporter molecule-labeled nucleotides by DNA polymerases. II. High-density labeling of natural DNA. *Nucleic Acids Research*, **31**, 2636–2646.
- 7 Kuwahara, M. *et al.* (2003) Simultaneous incorporation of three different modified nucleotides during PCR. *Nucleic Acids Research Supplement*, **3**, 37–38.
- 8 Kuwahara, M. *et al.* (2006) Systematic characterization of 2'-deoxynucleoside-5'-triphosphate analogs as substrates for DNA polymerases by polymerase chain reaction and kinetic studies on enzymatic production of modified DNA. *Nucleic Acids Research*, **34**, 5383–5394.
- 9 Bensimon, A. *et al.* (1994) Alignment and sensitive detection of DNA by moving interface. *Science*, **265**, 2096–2098.
- 10 Michalet, X. *et al.* (1997) Dynamic molecular combing: stretching the whole human genome for high-resolution studies. *Science*, **277**, 1518–1523.
- 11 Kearns, G.J. *et al.* (2006) Substrates for direct imaging of chemically functionalized SiO₂ surfaces by transmission electron microscopy. *Analytical Chemistry*, **78**, 298–303.
- 12 Dimalanta, E.T. *et al.* (2004) A microfluidic system for large DNA molecule arrays. *Analytical Chemistry*, **76**, 5293–5301.
- 13 Cheung, S.W. *et al.* (2005) Development and validation of a CGH microarray for

- clinical cytogenetic diagnosis. *Genetics in Medicine*, **7**, 422–432.
- 14 Stankiewicz, P. and Lupski, J.R. (2002) Molecular-evolutionary mechanisms for genomic disorders. *Current Opinion in Genetics & Development*, **12**, 312–319.
 - 15 Cheung, J. *et al.* (2003) Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biology*, **4**, R25.
 - 16 Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.
 - 17 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
 - 18 Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
 - 19 Chaisson, M. *et al.* (2004) Fragment assembly with short reads. *Bioinformatics*, **20**, 2067–2074.
 - 20 Whiteford, N. *et al.* (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*, **33**, e171.
 - 21 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science*, **302**, 1401–1404.
 - 22 Sebat, J. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **304**, 525–528.
 - 23 Tuzun, E. *et al.* (2005) Fine-scale structural variation of the human genome. *Nature Genetics*, **37**, 727–732.
 - 24 Korb, J.O. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
 - 25 Redon, R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
 - 26 Ng, P. *et al.* (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Research*, **34**, e84.
 - 27 Saha, S. *et al.* (2002) Using the transcriptome to annotate the genome. *Nature Biotechnology*, **20**, 508–512.
 - 28 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
 - 29 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
 - 30 Lockhart, D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675–1680.
 - 31 Bainbridge, M.N. *et al.* (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing by synthesis approach. *BMC Genomics*, **7**, 246–256.
 - 32 Cheung, F. *et al.* (2006) Sequencing *Medicago truncatula* expressed sequence tags using 454 Life Sciences technology. *BMC Genomics*, **7**, 272–281.
 - 33 Torres, T.T. *et al.* (2008) Gene expression profiling by massively parallel sequencing. *Genome Research*, **18**, 172–177.
 - 34 Morin, R.D. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research*, **18**, 610–621.
 - 35 Nei, M. and Li, W. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5269–5273.
 - 36 Lynch, M. (2007) *The Origins of Genome Architecture*. Sinauer Associates, Inc. Publishers, Sunderland, MA.

10

A Single DNA Molecule Barcoding Method with Applications in DNA Mapping and Molecular Haplotyping

Ming Xiao and Pui-Yan Kwok

10.1

Introduction

The success of the Human Genome Project (HGP) is largely due to the continuous development of Sanger sequencing method through parallelization, automation, miniaturization, better chemistry, and informatics. As the workhorse of the Human Genome Project, Sanger sequencing method has dominated the DNA sequencing field for nearly three decades, and its 800 Q20 base read length is still the gold standard, which no other sequencing methods can match [1]. However, its serial electrophoretic processing of DNA samples through the microchannels will not be able to compete with the newly emerging massively parallel sequencing technologies (e.g., ABI SOLiD, 454 FLX, Solexa Genome Analyzer) in further reducing the sequencing costs toward the goal of the \$1000 genome. These newly emerging sequencing technologies can be roughly grouped into two categories based on the detection methods, sequencing either by ensemble detection or by single-molecule detection. Since multiple DNA copies are needed in ensemble detection, the genetic information such as haplotype and RNA splicing pattern is lost during the process. Sequencing by single-molecule detection holds the great promise to recover the haplotype information; however, the read length of current single-molecule sequencing method (e.g., Helicos tSMS) is merely 50 bp or less, which is far shorter than the average distance of 1 kb between two SNPs. Just like their predecessor, Sanger sequencing method, the critical genetic information such as haplotypes and RNA splicing pattern is still difficult to obtain with these “next-generation” sequencing technologies. The future sequencing technology should resemble the DNA replication system in every cell, simply taking apart all the chromosomes, and reading each base of the chromosomes from end to end.

In this initial attempt toward our final goal of reading each base of a chromosome from end to end, we developed a single DNA molecule barcoding method that can obtain genetic information across genomic regions of over 10 kb. Our DNA barcoding strategy is based on direct fluorescent imaging and localization of multiple

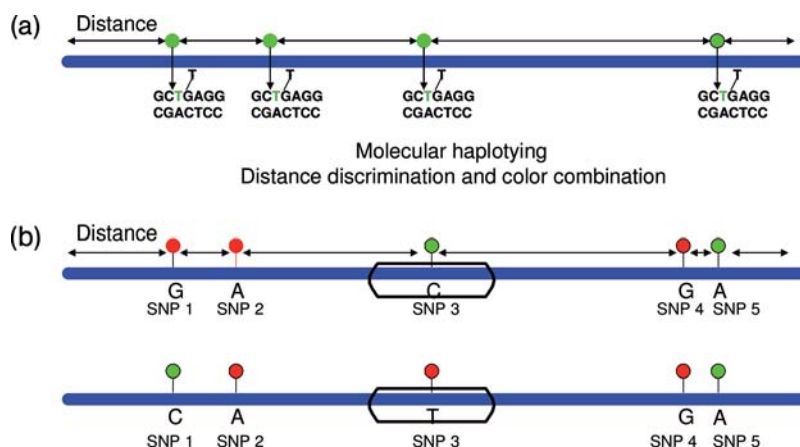


Figure 10.1 Single DNA molecule barcoding. (a) Sequence motif map with single DNA molecular mapping. Unique "sequence motif map" where each sequence motif (GCTGAGG) is recognized and nicked by a nicking endonuclease. After nicking, a Cy3 (green) fluorescent terminator is incorporated at the nicking site (green T) by polymerase, as the native T is displaced. The distance between labels along the YOYO-stained DNA backbone (blue) is determined by fluorescence single-molecule detection with TIRF microscopy. (b) Haplotype

barcodes. Unique "haplotype barcodes" for the two haplotypes where each allele of five SNPs is labeled with either a Cy5 (red) or a Cy3 (green) fluorescent padlock probe. Circularized padlock probes were used to label the alleles with fluorescent dyes. The distance and color combinations between labels along the YOYO-stained DNA backbone (blue) are determined by fluorescence single-molecule detection with TIRF microscopy. The haplotype can then be inferred from the "barcode."

sequence motifs or polymorphic sites on a DNA molecule. Individual genomic DNA molecules or long-range PCR fragments are labeled with fluorescent dyes at specific sequence motifs or polymorphic sites. The labeled DNA molecules are then stretched into linear form on a modified glass surface and imaged using total internal reflection fluorescence (TIRF) microscopy. By determining the positions and the colors of the fluorescent labels with respect to the DNA backbone, the distribution of the sequence motifs or polymorphic sites can be established with good accuracy, in a manner similar to reading a barcode [2]. We have successfully applied this single DNA molecule barcoding method in DNA mapping (Figure 10.1a) and molecular haplotyping (Figure 10.1b).

10.2

Critical Techniques in the Single DNA Molecule Barcoding Method

In the DNA barcoding method, one has to determine the positions and colors of the fluorescent labels found on the DNA molecule being analyzed. Five critical steps are involved: preparation of long DNA fragments, fluorescent tagging of specific sequences or polymorphic sites, stretching the DNA fragments into linear form on modified glass surface, multicolor TIRF imaging, and image analysis to localize the

individual fluorescent labels on the DNA and infer the genetic information. These techniques are interrelated, with the choice of one method in each step influencing and limiting the approaches one can take in the other steps.

The DNA barcoding method can work directly either with long genomic DNA fragments or with DNA fragments amplified by long-range PCR. Long-range PCR is a mature technique, with robust protocols and highly efficient enzymes commercially available for this purpose [3]. In fact, long-range PCR up to 17 kb was done routinely to amplify all the unique sequences in the human genome (<http://genome.perlegen.com/pcr/>).

Sequence-specific or allele-specific labeling genetic markers found on the long double-stranded DNA (dsDNA) molecules is probably the most challenging part of the method. In DNA mapping, individual genomic DNA molecules are labeled with fluorescent dyes at specific sequence motifs by the action of a nicking endonuclease followed by the incorporation of dye terminators with a DNA polymerase [4]. In molecular haplotyping, hybridization of sequence-specific probes with DNA molecules to form stable Watson–Crick duplexes is the only practicable approach for SNP recognition, since there are millions of SNPs in the genome. Moreover, the fluorescent tagging has to be allele specific, or no genetic information can be obtained. A variation of the padlock probe ligation approach [5] has been used in molecular haplotyping. This strategy involves three steps that are accomplished in a closed tube experiment. A long probe (75–100 bases) is designed and synthesized with the two ends containing DNA sequences complementary to the flanking sequences of an SNP, such that when the probe is hybridized to the target DNA, it forms an incomplete ring with a one-base gap at the polymorphic site. Deoxynucleoside triphosphates bearing a fluorescent label can then be used to fill the gap with the aid of DNA polymerase, forming a ring structure with a “nick” in the double-stranded probe–target complex. DNA ligase, already present in the reaction mixture, ligates the two ends of the padlock probe together to form a circular DNA probe that intertwines with the target DNA as the hybridized probe forms a double helix structure with the target DNA. This structure proves to be stable throughout the processes of DNA purification, linearization, and imaging.

DNA combing techniques (stretching long DNA molecules into linear form) have been widely used in genomic and genetic studies, such as optical mapping [6], fiber fluorescence *in situ* hybridization [7], genetic disease screening, and molecular diagnostics [8]. In these studies, the long DNA molecules are linearized in solution either as they flow through a microfluidic channel [9] or as they are stretched on a solid surface [10]. To linearize a piece of DNA on a solid surface, the end of a DNA molecule is first anchored to a hydrophobic surface (typically modified glass) by adsorption. We have developed a method of stretching dsDNA molecules as short as 4 kb on a polyelectrolyte multilayer (PEM)-modified glass surface [11]. The PEM-modified glass surface is constructed by sequential deposition of poly(allylamine) and poly(acrylic acid) with outermost layer being poly(allylamine) [12]. Since the outermost polymer layer is positively charged, the negatively charged DNA molecule is easily anchored and stretched on the modified glass surface. Moreover, an added advantage of the PEM-modified surface is that it has extremely low fluorescence

background, which is achieved by electrostatic repulsion of fluorescent impurities by the multilayer charged polymer [13]. The low fluorescence background surface makes it possible to detect single fluorescent dye molecules tagged along DNA backbone.

Individual fluorescent labels on the DNA backbone are imaged using multicolor TIRF microscopy, a technique capable of localizing single fluorescent dye molecules with nanometer-scale accuracy [14]. The TIRFM system was based on an Olympus IX-71 microscope with a custom-modified Olympus TIRFM Fiber Illuminator and a 100-SAPO objective. The current system is capable of detecting three colors. DNA backbone is stained with YOYO-1, which is excited by using 488-nm wide-field excitation from a mercury lamp. Cy3 (green) and Cy5 (red) fluorophores are used to label the sequence motifs or SNP polymorphic sites and are excited by using 543-nm and 628-nm helium–neon lasers, respectively. The spatial localization of individual dye molecules at the polymorphic sites or sequence motif sites is based on centroid analysis [15]. The centroid analysis strategy relies on the observation that a fluorescent molecule forms a diffraction-limited image of width $\lambda/2$, but the center of the distribution (which under appropriate conditions corresponds to the position of the dye) can be localized to arbitrarily high precision by fitting to a two-dimensional elliptical Gaussian point spread function (PSF), if a sufficient number of photons were collected.

Custom-written software in IDL (Research Systems, Inc., Boulder, CO, USA) was used for image analysis. The software can extract from the image DNA molecules based on the intensity of DNA backbone staining dye (YOYO-1) and single fluorescent dye labels. The extracted images are then merged and the individual dye labels are superimposed onto the DNA backbone. Sequence motif maps are then constructed and the haplotypes inferred after localizing the dye labels along the DNA backbone.

10.3

Single DNA Molecule Mapping

DNA mapping is an important analytical tool in genomic sequence assembly, medical diagnostics, and pathogen identification [16–18]. The current strategy for DNA mapping is based on sizing DNA fragments generated by enzymatic digestion of genomic DNA with restriction endonucleases. More recently, several linear DNA mapping techniques have been developed with the goal of mapping DNA in their native states. The DNA molecular combing and optical mapping techniques interrogate multiple sequence sites on single DNA molecules deposited on a glass surface, which has been used to detect disease-related mutations and map several microbial genomes [6, 8, 19]. A direct linear analysis (DLA) technique, in which a long dsDNA molecule was tagged at specific sequence sites with fluorescent dyes and stretched into linear form as it flowed through a microfluidic channel, has also been developed for DNA mapping [9]. These techniques not only provide the location of restriction or fluorescent labeling of sites, but also preserve the order of the restriction or fluorescent labeling of sites within the DNA molecule.

Our DNA mapping method is based on direct imaging of individual DNA molecules and localization of multiple sequence motifs on these molecules.

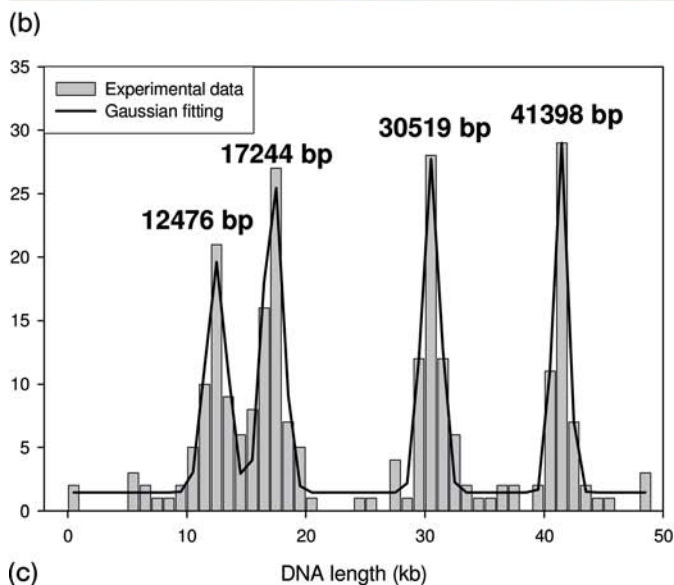
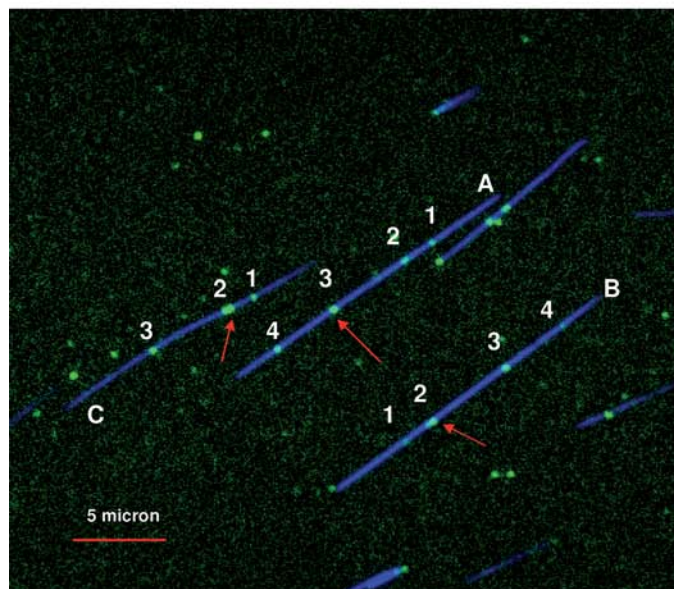
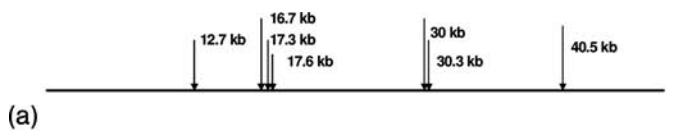
Individual genomic DNA molecules are labeled with fluorescent dyes at specific sequence motifs. The sequence-specific labeling starts with introducing nicks in dsDNA at specific sequence motifs recognized by nicking endonucleases, which cleave only one strand of a dsDNA substrate [20]. DNA polymerase then incorporates fluorescent dye terminators at these nicking sites (Figure 10.1a). Currently, there are six commercially available nicking endonucleases (New England Biolabs) with recognition sequence motifs ranging from three to seven bases long.

The labeled DNA molecules are stretched into linear form and imaged on a modified glass surface. The distribution of the sequence motifs recognized by the nicking endonuclease can be established with great accuracy. With this approach, we constructed sequence motif maps of the lambda phage, a strain of human adenovirus and several strains of human rhinoviruses. Because of the simplicity of this mapping strategy (single DNA molecule analysis, high accuracy, and potential of high throughput), it will likely find applications in DNA mapping, medical diagnostics, and especially in rapid identification of microbial pathogens.

10.3.1

Sequence Motif Maps of Lambda DNA

Lambda DNA is used as a model system to construct a sequence motif map. Figure 10.2a shows the distribution of its seven nick endonuclease Nb.BbvC I recognition sites. The solid black line represents the backbone of the lambda DNA and the black arrow indicates the positions of the predicted Nb.BbvC I sites. Two images were taken and superimposed to produce a composite picture of the DNA molecules. Figure 10.2b is a false-color two-channel composite image showing the stretched DNA contours (YOYO in blue) and labeled sites (Tamra-ddUTP in green). Three DNA molecules are nearly fully stretched (A, B, and C) with contour lengths of 19.8, 19.5, and 16.9 μm , respectively. Although the data suggest that DNA molecules A and B are overstretched at 0.41 nm/bp, compared to the solution conformation of 0.34 nm/bp, this may be due to the effect of YOYO staining [21]. The rest of the DNA molecules are either broken or folded back onto themselves, giving lengths much shorter than that predicted. There are also occasional Tamra dye signals (green) not associated with the DNA backbone. These are most likely the result of either fluorescent impurities on the coverslip or free Tamra-ddUTP. DNA fragments A and B in Figure 10.2b have four Tamra labels (green) along the DNA backbone, and DNA fragment C has three green labels. The signal for two of the green labels (red arrows) is much stronger and occupies more pixels than that corresponding to a single fluorescent dye, indicating that several green labels have clustered together and cannot be resolved due to light diffraction limits of the instrument. The two clusters most likely correspond to the predicted sites 2, 3, and 4 and sites 5 and 6, as they are separated by no more than 1000 bp. Accordingly, the seven Nb.BbvC I sites of lambda DNA are collapsed to four resolvable sites, with the middle two signals stronger than the outer two signals. The distances between the labels were calculated with respect to the DNA backbone starting from the top right end of the DNA backbone. The positions of four green labels on DNA molecule A starting from the



right end are 12.6, 17.9, 31.0, and 40.3 kb, respectively, but they are at 6.8, 17.4, 31.0, and 35.8 kb for DNA molecule B. Clearly, the two DNA molecules were in opposite orientation. DNA molecules that fulfill the following two criteria were selected for analysis. First, the labeled DNA fragments must be nearly fully stretched (longer than 15 μm for lambda DNA). Second, DNA fragments must have at least three labels so that the relative distances between the labels can be used to establish the orientation of the DNA molecules. Figure 10.2c is the sequence motif map of lambda DNA based on the analysis of 81 molecules that met the criteria we set. Four peaks were calculated as 12 476, 17 244, 30 519, and 41 398 bp from one end, in good agreement with the predicted distribution of sequence motif. The closest two peaks are 5 kb and they are clearly distinguishable. More than 70% of the fully stretched DNA molecules have more than three labels, indicating that the labeling efficiency is relatively high.

10.3.2

Identification of Several Viral Genomes

The sequence motif maps of human adenovirus and rhinovirus genomes were constructed to demonstrate the capability of our approach to rapidly map and identify pathogen genomes. The genome of human adenovirus type 2 is 35.5 kb in length, and the distribution of Nb.BbvC I sites is shown in Figure 10.3a. The genome contains nine Nb.BbvC I sites, of which seven are resolvable. Figure 10.3b is a false-color two-channel composite image showing several fully stretched DNA molecules (YOYO in blue) with all seven resolvable Nb.BbvC I sites labeled with Tamra-ddUTP (green). A total of 105 DNA molecules were used in constructing the sequence motif map shown in the bottom graph of Figure 10.3c. The seven peaks were found at 2.8, 10.9, 15.4, 18.7, 24.7, 31.3, and 34.4 kb from one end, compared to the expected positions of 3.3, 11.1, 15.2, 18.2, 23.9, 30.1, and 33.3 kb. Based on previously published phylogenetic analyses of the human adenovirus genome sequences [21, 22], virtual sequence motif maps of Nb.BbvC I sites of several human adenovirus genomes were constructed. Clearly, the Nb.BbvC I maps of six major types (A–F) of human adenoviruses are quite different. Even the strains of closely related subtype can be quite easily distinguished. Therefore, one can distinguish the different viral strains by just comparing the map obtained experimentally with the known virtual sequence motif maps.

Human rhinovirus is an RNA virus and the length of the genomes of different types of rhinovirus is about 7.2 kb. Before constructing its Nb.Bsm I map, nearly

Figure 10.2 Sequence motif map of lambda DNA. (a) The predicted Nb.BbvC I map of lambda DNA. Positions of the nicking sites are indicated by arrows. Nicking sites 2–4 and 5–6 are closely clustered and are not resolvable due to the limits of optical diffraction. (b) In the intensity scaled composite image of linear lambda DNA, the Nb.BbvC I sites (labeled with Tamra-ddUTP) are shown as green spots and the DNA backbone (labeled with YOYO) is shown as blue lines. Owing to the diffraction limits of the

microscope, only four labels can be fully resolved. In this field, two DNA fragments (A and B) are fully labeled while one fragment (C) has three labels. Red arrows point to clustered sites, some of them are brighter than other because of the presence of multiple labels. (c) The sequence motif map in the bottom graph was obtained by analyzing 61 single-molecule fluorescence images. The solid line is the Gaussian curve fitting and the peaks correspond well to the predicted locations of the sequence motif.

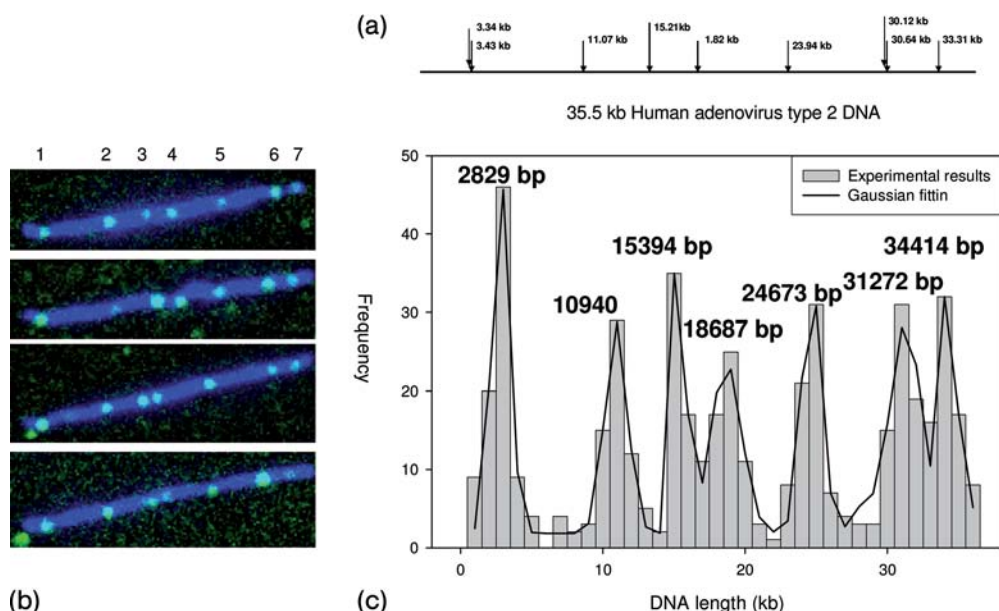


Figure 10.3 Sequence motif map of human adenovirus type 2. (a) Predicted Nb.BbvC I map of human adenovirus type 2. Nine sites are found on the 35.5 kb viral DNA with two sets of clustered sites (1–2 and 7–8), leading to seven resolvable labels. (b) In the intensity scaled composite images of four fully labeled human adenovirus type 2 DNA, the Nb.BbvC I sites (labeled with Tamra-ddUTP) are shown as green spots and the DNA backbone (labeled with

YOYO) is shown as blue lines. Owing to the diffraction limits of the microscope, only seven labels can be fully resolved. Labels 1 and 6 are generally brighter than the other labels due to clustering. (c) The sequence motif map in the graph was obtained by analyzing 63 single-molecule fluorescence images. The solid line is the Gaussian curve fitting and the peaks correspond well to the predicted locations of the sequence motif.

full-length dsDNAs (6.4 of 7.2 kb) were generated by reverse transcription followed by PCR. To see if one could use this approach to identify viral isolates, we conducted a set of studies in which the identities of the viral strains were unknown to those performing the experiments and data analysis. Four anonymous strains of human rhinoviruses were obtained from our collaborators. Full-length dsDNA was generated by reverse transcription and long-range PCR as before, using conserved sequences as PCR primers. After DNA nicking with Nb.Bsm I (GCATTC) and labeling with Tamra-ddCTP, the optical maps of the four viral strains were obtained and analyzed. The strains were identified as HRV15, HRV28, HRV36, and HRV73 by comparing the predicted sequence motif maps with the constructed maps (data not shown).

10.4

Molecular Haplotyping

The power to detect association in a genetic study is greatly enhanced when one compares the haplotypes of the cases and controls rather than working exclusively

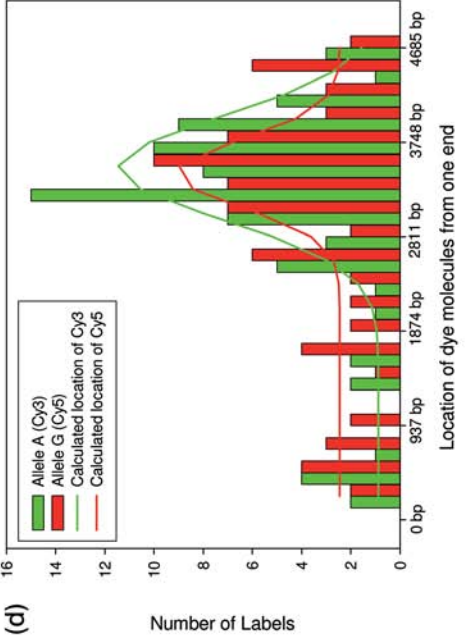
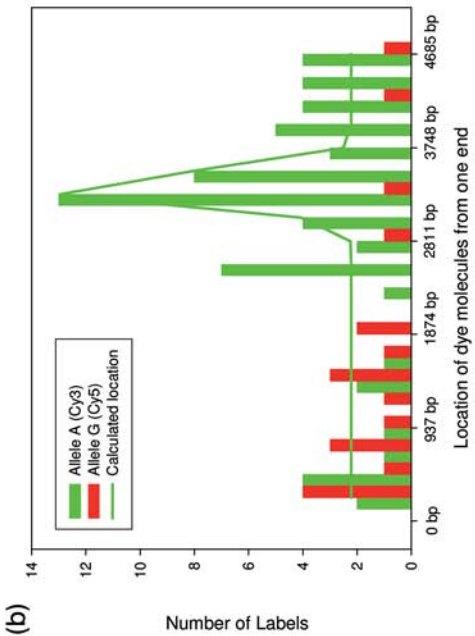
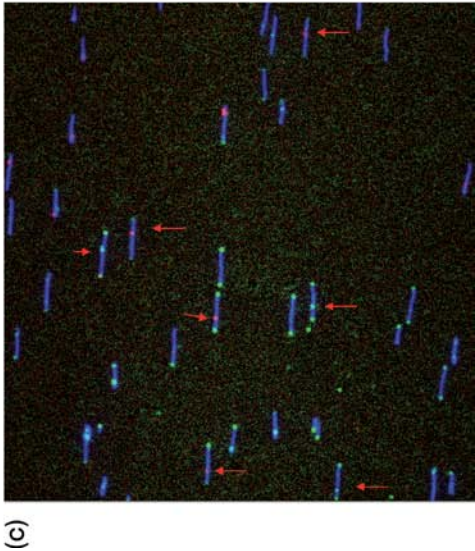
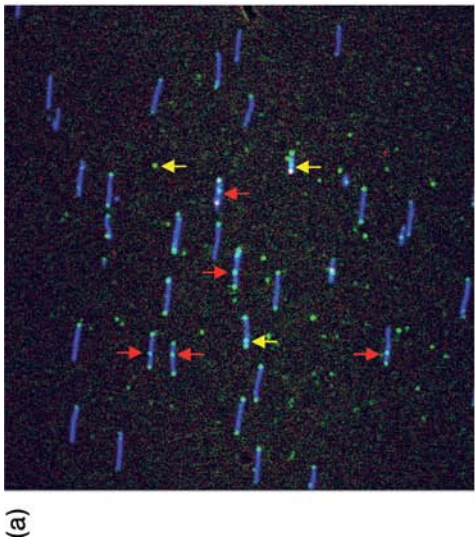
with SNPs [23]. However, determining the haplotypes in a diploid individual is a major technical challenge. A reliable, accurate, and high-throughput molecular haplotyping technology is urgently needed.

Our single-molecule barcoding method is used to determine haplotypes by directly imaging multiple polymorphic sites on individual DNA molecules simultaneously. The method starts with long-range PCR amplification of target DNA segments containing the polymorphic sites, followed by allele-specific labeling of polymorphic alleles with fluorescent dye molecules, imaging the linearly stretched single DNA molecules, and determining the nature and positions of the fluorescent dyes along the DNA molecules. By determining the colors and positions of the fluorescent labels with respect to the backbone at polymorphic sites, the haplotype may be inferred with great accuracy, even when the DNA fragments are not fully labeled (Figure 10.1b). The feasibility of this approach is demonstrated by the determination of the haplotypes of a 9.3-kb DNA fragment containing four SNPs in a region on human chromosome 17 that is linked to the susceptibility of the skin disease psoriasis [24].

10.4.1

Localization of Polymorphic Alleles Tagged by Single Fluorescent Dye Molecules Along DNA Backbones

Since the haplotype barcodes consist of allele-specific color tagging and distance discrimination between labels, in the first set of experiments, we sought to show that padlock probe labeling of SNP is allele specific and the labeled SNP can be accurately localized along the DNA backbone. To aid with the distance measurements, we used Cy3-labeled PCR primers to amplify a 9.3-kb fragment containing the SNP rs12797, a G > A polymorphism. Using the gap-filled ligation approach, the two alleles were tagged with Cy3-dATP (green) and Cy5-dGTP (red). The DNA backbone was stained with YOYO (blue). Three images (with the green, red, and blue channels) were taken and superimposed to produce a composite picture of the DNA molecules. Figure 10.3a is a false-color three-channel composite image showing the stretched DNA contours and allelic labels (with Cy5-dGTP in red, Cy3-dATP in green, and YOYO in blue). About 30 DNA molecules are shown in this image and 20 of them are fully stretched, with a mean contour length of 3.5 μm . This suggests slight overstretching of the DNA of 0.38 nm/bp. Most of the DNA fragments in Figure 10.4a have Cy3 dyes at both ends, and some of them have a Cy3 in the middle (as shown by the red arrows), indicating the presence of Cy3-labeled probe on the backbone. The Cy3 label (A allele) was calculated to be at position 3311 bp, which is in excellent agreement with the expected position of 3291 bp from one end (Figure 10.4b). However, few red labels (G allele) were detected, and these were distributed randomly, confirming the fact that this DNA sample is A > A homozygote for SNP rs12797. Figure 10.4c and d shows the results of another experiment in which the DNA sample from an rs12797 G > A heterozygote was labeled and the distances measured. In this case, both green and red labels (A and G alleles) were detected at about 3459 ± 492 and 3413 ± 372 bp from one end, respectively, compared to the expected position of 3291 bp from one end. The proportion of



red labels and green labels found on the DNA backbone is roughly 50 : 50, as expected from a heterozygous sample.

10.4.2

Direct Haplotype Determination of a Human DNA Sample

We demonstrated this technology's ability to correctly determine a haplotype consisting of four SNPs. Once again, we studied the 9.3-kb DNA segment of human chromosome 17, containing markers rs878906(C > T) (SNP 3-1), rs12797 (G > A) (SNP 3-2), rs734232(G > A) (SNP 3-3), and rs745318(C > T) (SNP 3-4). As before, the alleles were tagged with gap-filled padlock probes. In this case, the G and C alleles were labeled with red Cy5-dGTP and Cy5-dCTP, and the A and T alleles were tagged with Cy3-dATP and Cy3-dUTP, respectively. An additional green-channel dye was introduced at one end during long-range PCR by using a Cy3-labeled primer. This end label was used to indicate the orientation of DNA molecules. The relative distance between polymorphic sites starting from the end label is shown in Figure 10.5a. Figure 10.5b is a false-color composite image of all three channels from a typical experiment with a DNA sample from an individual who is heterozygous at all four SNPs. Most of the DNA fragments are fully stretched, and some fully stretched DNA molecules show more than one internal labels. As the current labeling efficiency is about 25% for each SNP, one should find an average of four DNA fragments out of 1000 DNA molecules with all four polymorphic sites labeled. Considering the fact that about 40% of the DNA fragments are fully stretched and are therefore suitable for analysis, 2500 DNA molecules must be scanned to find one fully labeled DNA fragment. However, because the spatial localization of fluorescent dyes is very accurate, some partially labeled DNA fragments can be used to assemble the haplotype, as long as they fulfill three criteria. First, the labeled DNA fragments must be fully stretched so that the label positions may be accurately determined. Second, they must have an end label, to allow them to be oriented and aligned

Figure 10.4 Localization of fluorescently labeled alleles on dsDNA backbone. (a) An intensity scaled composite image of all three channels. The alleles of the SNP rs12797 were labeled with Cy3 dye (green) for the A allele and Cy5 dye (red) for the G allele. The positions of labeled alleles are indicated with red arrow. Few red labels were observed, indicating this sample is AA homozygous. Yellow arrows indicate dyes at incorrect positions. (b) Histogram of the distance distribution of the results from (a). Red bars indicate the G allele and green bars represent the A allele, respectively. The Gaussian curve fitting shows a green peak at 3311 ± 161 bp from one end, which is consistent with the expected distance of 3291 bp. Eighty-six molecules were examined, 66 with Cy3 internal labels and 20 with Cy5 internal labels were observed in total. (c) An intensity scaled composite image of all three channels. The alleles of the SNP rs12797 were labeled with Cy3 (green) for the A allele and Cy5 (red) for the G allele. The positions of labeled alleles are indicated with red arrows. Both Cy3 and Cy5 labels were observed, indicating this sample is GA heterozygous. (d) Histogram of the distance distribution of the results from (c). Red indicates the G allele and green represents the A allele. The Gaussian curve fitting shows a green peak and a red peak at 3459 ± 492 and 3413 ± 372 bp from one end, respectively, which is consistent with the actual distance of 3291 bp. A total of 228 DNA molecules were examined, from which 73 Cy3 labels and 69 Cy5 labels were analyzed.

properly. Third, DNA fragments must have at least two polymorphic sites labeled to show the haplotype relationship between them. One such DNA molecule with two internal labels is shown in the uppermost inset of Figure 10.5b.

Once the label positions have been determined, each observed label is then matched to a known locus. Briefly, for a given fragment, we generate all possible locus–label matchings, consistent with maintaining the observed linear order of the labels. Each of these matchings is assigned a score $S = \sum_{i=1}^N e^{-d_i^2/\sigma^2}$, where d_i is the distance between label i and its assigned locus, N is the number of labels, and Σ is based on the observed standard deviation for label position measurements (approximately 5–10%). The matching with the highest score is selected. In this case, they were determined to be the alleles of SNP 3-3 and SNP 3-4, with the alleles being G (SNP 3-3)-C(SNP 3-4). Another fragment, with three internal labels, are shown in the lower inset, with the allele labels being C(SNP 3-1)-A(SNP 3-2)-T(SNP 3-4). This score is then multiplied by the number of labels, reflecting the increased confidence in the locus–label matching that comes from having multiple labels present. A running score is kept for all possible haplotypes, and the score for the fragment is added to the score for the appropriate haplotypes. If a fragment has fewer labels than there are loci,

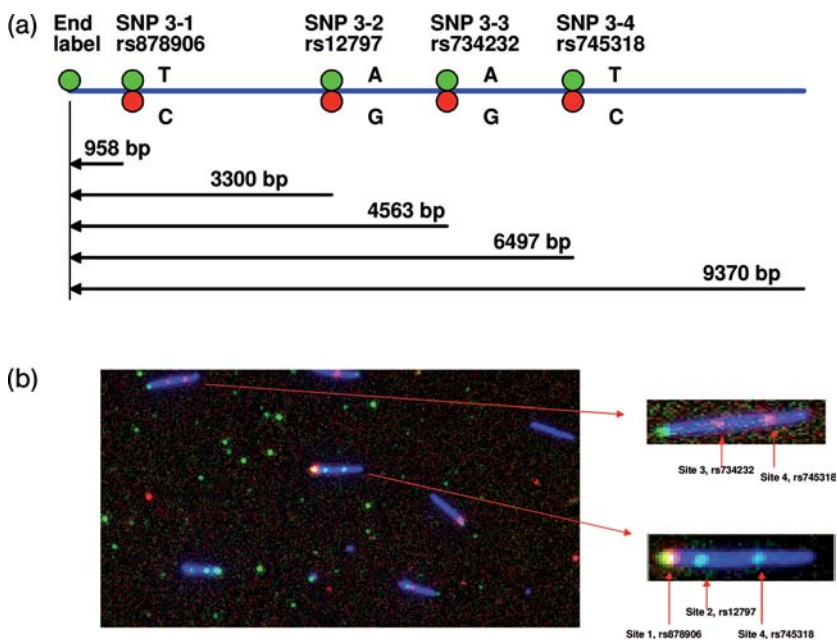


Figure 10.5 Haplotype barcode. (a) Relative locations of the polymorphic sites, their alleles, and the labels assigned to each allele. Green represents Cy3 and red represents Cy5. (b) Rescaled false-color composite image of all three channels, showing DNA fragments with multiple labels, which have been identified and tagged based on their position on the DNA fragment.

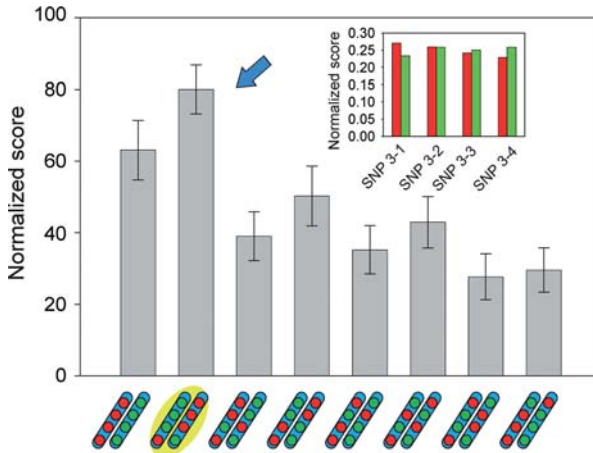


Figure 10.6 All eight possible heterozygous haplotypes with their scores. The arrow indicates the score of the highlighted haplotype, RGGG/GRRR. *Inset:* Scores for Cy3 and Cy5 at each individual locus, confirming that all four loci are heterozygous.

it may correspond to several possible haplotypes. In this case, the scores for all possible matching haplotypes are incremented. This way, partially labeled fragments can contribute to the calculation of the haplotype. One can also take advantage of the fact that a diploid sample with four heterozygous SNPs can contain only two distinct haplotypes from eight pairs of complementary haplotypes and construct haplotypes with partially labeled DNA molecules. The horizontal axis of Figure 10.6 shows all eight possible pairs of haplotypes, where each pair consists of two complementary haplotypes, with each allele represented by a color, either red or green. Figure 10.6 shows the results of scoring 72 doubly or triply labeled, well-stretched DNA fragments with end labels. Of those, 77% had two labels, 21% had three labels, and 2% had four labels. The top-scored haplotype is red-green-green-green/green-red-red-red (RGGG/GRRR), corresponding to either T-G-G-C or C-A-A-T for the loci rs878906, rs12797, rs734232, and rs745318, respectively. The score for this haplotype pair is more than 30% higher than the next highest scored haplotype pair, clearly indicating that this is the haplotype pair observed for this sample. This result was confirmed by parental genotyping of all four SNPs (data not shown). The inset shows the normalized score assigned to each of the four positions corresponding to the number of times a Cy3 or Cy5 was seen at that position. It confirms that all four positions are heterozygous, because all four positions show instances of both Cy3 and Cy5.

10.5 Discussion

As evident from the results of DNA mapping and molecule haplotyping discussed in this chapter, the single DNA molecule barcoding method is very versatile and can be

applied in a wide range of genetic and genomic applications. For example, one can barcode alternative RNA splicing patterns and record the transcripts' full profiles (in terms of varieties and quantities); this technique can also be used to study genomic structural variations, such as barcoding copy number variations. Among the steps of our single DNA molecule barcoding method, fluorescent labeling of specific sequences and polymorphic sites is the most critical step and it largely determines the usefulness and versatility of the method.

Most methods of labeling specific DNA sequence on dsDNA molecules are based on noncovalent binding, and the sequence recognition depends on the relative binding affinity of the probe. At the single-molecule level, the dissociation of the labeled probe is significant if the unbound probes have to be removed, thus significantly reducing the labeling efficiency. In our nick-labeling scheme, the specificity is determined by both the enzymatic nicking reaction and the fluorescent nucleotide incorporation reaction. Furthermore, the single fluorescent dye molecules are covalently bound to the dsDNA molecules and are therefore not subjected to the variation of binding constants. Similarly, the padlock probe used in molecule haplotyping is topologically bound to the dsDNA target and proves to be very stable during subsequent purification and imaging process. These properties make our approach superior to those that are currently in use.

Since the localization of fluorescent dye molecules takes place with respect to the DNA backbone, the degree of the DNA stretching directly affects the results of DNA barcoding. There are a number of ways to improve on our mounting and stretching DNA molecules on modified glass surface. One promising method is to elongate the dsDNA molecules in nanometer-size channels [25, 26], which can be made as small as 5 nm in width. The confinement of elongated DNA molecules not only provides uniform DNA stretching, but can also be easily integrated into a fully automated imaging system.

Our current single fluorescence imaging system is based on a three-color system: a blue channel (YOYO-1) for the DNA backbone; a green (Cy3) and a red (Cy5) channel for the single dye molecule detection. In sequence motif mapping, two different sequence motifs can be tagged with different color dyes, and this way, more flexible and finer maps can be constructed. Of course, a five-color system is ideal with a different color matching each of the four DNA bases. The current resolution of localization of single dye molecule is limited by the diffraction limit on the order of 250 nm or about 800 bp. This means that two polymorphic sites or sequence motif sites under interrogation have to be at least 800 bp apart to be resolved as separate sites. This should be adequate for most DNA barcoding applications. If higher resolution is needed, there are methods by which two dye molecules of the same color can be resolved down to 10 nm [27].

In conclusion, the nick-labeling scheme improved the chemistry of optical mapping and enabled us to map sequence motifs along long double-stranded DNA molecules. This technology provides the linear ordered map of sequence motifs, which cannot be obtained directly with gel-based restriction mapping. As less than 100 molecules are needed to construct the sequence motif map in our approach, the drastic reduction of DNA material used in the labeling step may be possible, once the

labeling procedures are miniaturized. By comparing the sequence motif map of a microbial isolate with the predicted (virtual) map in the database, one can determine the identity of a microbe accurately and efficiently. We also provide a proof-in-principle study of obtaining haplotype information directly with our DNA barcoding method. Once the labeling efficiency is further improved and the system automated, this approach can be used to determine haplotypes accurately, quickly, and at low cost and can lead to a practical molecular haplotyping technique suitable for the average laboratory.

References

- 1 Metzker, M.L. (2005) Emerging technologies in DNA sequencing. *Genome Research*, **15**, 1767–1776.
- 2 Kwok, P.Y. and Xiao, M. (2004) Single-molecule analysis for molecular haplotyping. *Human Mutation*, **23**, 442–446.
- 3 Barnes, W.M. (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 2216–2220.
- 4 Xiao, M., Phong, A., Ha, C., Chan, T.F., Cai, D.M., Leung, L., Wan, E., Kistler, A.L., DeRisi, J.L., Selvin, P.R. and Kwok, P.Y. (2007) Rapid DNA mapping by fluorescent single molecule detection. *Nucleic Acids Research*, **35**, e16.
- 5 Nilsson, M., Malmgren, H., Samiotaki, M., Kwiatkowski, M., Chowdhary, B.P. and Landegren, U. (1994) Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science*, **265**, 2085–2088.
- 6 Schwartz, D.C., Li, X., Hernandez, L.I., Ramnarain, S.P., Huff, E.J. and Wang, Y.K. (1993) Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science*, **262**, 110–114.
- 7 Wiegant, J., Kalle, W., Mullenders, L., Brookes, S., Hoovers, J.M., Dauwerse, J.G., van Ommen, G.J. and Raap, A.K. (1992) High-resolution *in situ* hybridization using DNA halo preparations. *Human Molecular Genetics*, **1**, 587–591.
- 8 Gad, S., Klinger, M., Caux-Moncoutier, V., Pages-Berhouet, S., Gauthier-Villars, M., Coupier, I., Bensimon, A., Aurias, A. and Stoppa-Lyonnet, D. (2002) Bar code screening on combed DNA for large rearrangements of the BRCA1 and BRCA2 genes in French breast cancer families. *Journal of Medical Genetics*, **39**, 817–821.
- 9 Chan, E.Y., Goncalves, N.M., Haeusler, R.A., Hatch, A.J., Larson, J.W., Maletta, A.M., Yantz, G.R., Carstea, E.D., Fuchs, M., Wong, G.G., Gullans, S.R. and Gilmanshin, R. (2004) DNA mapping using microfluidic stretching and single-molecule detection of fluorescent site-specific tags. *Genome Research*, **14**, 1137–1146.
- 10 Michalet, X., Ekong, R., Fougereousse, F., Rousseaux, S., Schurra, C., Hornigold, N., van Slegtenhorst, M., Wolfe, J., Povey, S., Beckmann, J.S. and Bensimon, A. (1997) Dynamic molecular combing: stretching the whole human genome for high-resolution studies. *Science*, **277**, 1518–1523.
- 11 Chan, T.F., Ha, C., Phong, A., Cai, D., Wan, E., Leung, L., Kwok, P.Y. and Xiao, M. (2006) A simple DNA stretching method for fluorescence imaging of single DNA molecules. *Nucleic Acids Research*, **34**, e113.
- 12 Lvov, Y., Decher, G. and Sukhorukov, G. (1993) Assembly of thin films by means of successive deposition of alternate layers of

- DNA and poly(allylamine). *Macromolecules*, **26**, 5396–5399.
- 13 Kartalov, E.P., Unger, M.A. and Quake, S.R. (2003) Polyelectrolyte surface interface for single-molecule fluorescence studies of DNA polymerase. *BioTechniques*, **34**, 505–510.
 - 14 Yildiz, A., Forkey, J.N., McKinney, S.A., Ha, T., Goldman, Y.E. and Selvin, P.R. (2003) Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. *Science*, **300**, 2061–2065.
 - 15 Thompson, R.E., Larson, D.R. and Webb, W.W. (2002) Precise nanometer localization analysis for individual fluorescent probes. *Biophysical Journal*, **82**, 2775–2783.
 - 16 Ried, T., Liyanage, M., du Manoir, S., Heselmeyer, K., Auer, G., Macville, M. and Schrock, E. (1997) Tumor cytogenetics revisited: comparative genomic hybridization and spectral karyotyping. *Journal of Molecular Medicine*, **75**, 801–814.
 - 17 van Belkum, A. (1994) DNA fingerprinting of medically important microorganisms by use of PCR. *Clinical Microbiology Reviews*, **7**, 174–184.
 - 18 Wong, G.K., Yu, J., Thayer, E.C. and Olson, M.V. (1997) Multiple-complete-digest restriction fragment mapping: generating sequence-ready maps for large-scale DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 5225–5230.
 - 19 Reslewic, S., Zhou, S., Place, M., Zhang, Y., Briska, A., Goldstein, S., Churas, C., Runnheim, R., Forrest, D., Lim, A., Lapidus, A., Han, C.S., Roberts, G.P. and Schwartz, D.C. (2005) Whole-genome shotgun optical mapping of *Rhodospirillum rubrum*. *Applied and Environmental Microbiology*, **71**, 5511–5522.
 - 20 Morgan, R.D., Calvet, C., Demeter, M., Agra, R. and Kong, H. (2000) Characterization of the specific DNA nicking activity of restriction endonuclease N.BstNBI. *Biological Chemistry*, **381**, 1123–1125.
 - 21 Allard, A., Albinsson, B. and Wadell, G. (2001) Rapid typing of human adenoviruses by a general PCR combined with restriction endonuclease analysis. *Journal of Clinical Microbiology*, **39**, 498–505.
 - 22 Casas, I., Avellon, A., Mosquera, M., Jabado, O., Echevarria, J.E., Campos, R.H., Rewers, M., Perez-Brena, P., Lipkin, W.I. and Palacios, G. (2005) Molecular identification of adenoviruses in clinical samples by analyzing a partial hexon genomic region. *Journal of Clinical Microbiology*, **43**, 6176–6182.
 - 23 Douglas, J.A., Boehnke, M., Gillanders, E., Trent, J.M. and Gruber, S.B. (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nature Genetics*, **28**, 361–364.
 - 24 Helms, C., Cao, L., Krueger, J.G., Wijsman, E.M., Chamian, F., Gordon, D., Heffernan, M., Daw, J.A.W., Robarge, J., Ott, J., Kwok, P.Y., Menter, A. and Bowcock, A.M. (2003) A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis. *Nature Genetics*, **35**, 349–356.
 - 25 Reisner, W., Morton, K.J., Riehn, R., Wang, Y.M., Yu, Z., Rosen, M., Sturm, J.C., Chou, S.Y., Frey, E. and Austin, R.H. (2005) Statics and dynamics of single DNA molecules confined in nanochannels. *Physical Review Letters*, **94**, 196101.
 - 26 Tegenfeldt, J.O., Prinz, C., Cao, H., Chou, S., Reisner, W.W., Riehn, R., Wang, Y.M., Cox, E.C., Sturm, J.C., Silberzan, P. and Austin, R.H. (2004) From the cover: the dynamics of genomic-length DNA molecules in 100-nm channels. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 10979–10983.
 - 27 Gordon, M.P., Ha, T. and Selvin, P.R. (2004) Single-molecule high-resolution imaging with photobleaching. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 6462–6465.

11

Optical Sequencing: Acquisition from Mapped Single-Molecule Templates

Shiguo Zhou, Louise Pape, and David C. Schwartz

11.1

Introduction

Knowledge of nucleic acid sequence forms the basis for modern biological investigation as represented by the sequencing of thousands of genomes, which has ushered in the current “genomics era.” This first step in the “New Biology [1]” was taken by the Sanger capillary electrophoresis sequencing approach that has framed the current need for new technologies for commoditization of sequence information through dramatic cost reductions. In this regard, a collection of new sequencing technologies has emerged, obviating traditional clone library construction through the use of single-molecule analytes either through direct measurement or by *in situ* template amplification prior to sequence acquisition. In short, these developments are largely based on schemes employing sequencing-by-synthesis (SBS) approaches including Illumina Genome Analyzer platform (Illumina, Inc.) [2], polony sequencing [3, 4], pyrosequencing [5, 6] (Biotage, Inc., pyrosequencing platform; Roche-454 Life Sciences, Inc., GS FLX platform, using massively parallel picotiter plates [7]), single-pair fluorescent resonance energy transfer (spFRET) single-molecule arrays [8] (Helicos, Inc.), zero-mode waveguide sequencing [9], and optical sequencing [10, 11]. An early, somewhat complementary approach not employing SBS – massively parallel signature sequencing (MPSS) – obtained 16–20 bp signatures using cycles of enzymatic cleavage with a type II restriction endonuclease, adapter ligation, and hybridization of specialized probes [12]. Although commercialized versions of these emerging sequencing platforms, shown above in parentheses, offer substantially higher throughput and lower costs compared to traditional Sanger sequencing approaches, they do not necessarily portend commoditization of DNA sequence information fostering ubiquitous analysis of human populations as might be required for everyday clinical diagnostics. This is, in part, because many of these new platforms wisely trade overall throughput for massive acquisition of attenuated sequence read lengths (5–200 bp) – a bold and effective compromise for many applications, for example, expression profiling – cost-effectively supplanting

the use of hybridization chips by low-noise sequence signatures [2]. However, short-sequence reads complicate effective *de novo* sequencing efforts and make access to the repeat-rich portion of the human genome, often associated with disease, somewhat opaque to detailed analysis. In addition to these concerns remains the fact that costs (>\$100 000/human genome) are still too high for the large-scale analysis of human populations. As such, development of cost-effective approaches for comprehensive resequencing of human genomes remains a difficult challenge.

Platform designs engendering low-cost, high-information content sequence analysis can be envisioned using very large (~500 kb) genomic DNA molecules. This developmental tact is a radical departure from current “next-generation” sequencing platforms that solely analyze short templates less than 1 kb, as it offers many significant advantages:

- (i) Very large DNA molecules can be “barcoded” so that their genomic origin is unambiguously determined. In this way, random template molecules are identified and placed within the human genome so that acquired sequence reads, carried on such patterned molecules, have a well-defined genomic context (the nucleotide positions spanned by any given molecule are known within a sequenced genome). Importantly, this barcoding step allows a confident placement of short reads within dispersed sequence repeats, large gene families, and constitutively heterochromatic genomic regions located near centromeres and telomeres, whereas conventional sequence alignment approaches would suffer ambiguous multiple placements of reads 30 bp in length. In short, specialized algorithms and software have been developed that “match” or align single-molecule barcodes with a reference genome (e.g., hg17) [13, 14]. Accordingly, systems developed that perform such operations include optical mapping [15] and “nanocoding” [16]; these systems also offer routes for discovery and characterization of large-scale rearrangements associated with cancer genomes.
- (ii) Multiple sequence reads acquired from the same large template molecule offer information-rich data that easily span complex genomic rearrangements, simplify *de novo* assembly, and are intrinsically free from any assembly errors. Such attributes would simplify sequence analysis and lower associated costs.
- (iii) The combination of molecular barcoding of large DNA molecular templates and sequence acquisition simultaneously provides both fine- and coarse-scale genomic analyses. This concept may provide an economical route for a comprehensive and revealing genome analysis enabling tabulation or characterization of most human genomic alterations presented as polymorphisms or mutations. Consequently, having a finely “barcoded” and partly sequenced genome may provide just the essential analysis required by future applications focusing on whole genome diagnostics or personal genomics.

The optical sequencing platform [10, 11] was developed to embody the advantages of very large DNA molecule templates and is the focus of this chapter. Briefly, the essential components of optical sequencing center on the use of purely single-molecule SBS acquisition from very large, barcoded molecules. These large

double-stranded DNA molecules are unraveled, elongated, and arrayed on surfaces for barcoding and preparation as competent templates for sequence acquisition. Template preparation entails random nicking with DNase I, followed by creation of single-stranded gaps distributed within native molecules using T7 exonuclease. Cycles of polymerase-mediated incorporation of fluorochrome-labeled nucleotides, tracked by fluorescence microscopy, create “histories” of labeled nucleotide additions at gapped locations residing on individual molecules. The tabulation and analysis of these fluorescent punctates on barcoded molecules produce strings of sequence or “reads” (the punctates – the sites of labeled fluorochrome addition(s) – are below the spatial resolution of light microscopy). Another unique attribute of optical sequencing is that multiple additions are economically tabulated per cycle; in other words, during a “G-nucleotide” incorporation cycle, multiple labeled Gs are added and tabulated or counted when there are multiple “C-nucleotides” on the DNA template. Accordingly, a photobleaching step within each cycle is included for “resetting” the counter, allowing single-nucleotide sensitivity through diminution of fluorescence after fluorochrome counts are tabulated across all “competent” punctates, cycle after cycle. What follows is an overview of the interlocking steps (Figure 11.1) enabling optical sequencing.

11.2

The Optical Sequencing Cycle

- Step 1. DNA presentation, barcoding. A microfluidic device unravels and arrays native genomic DNA molecules in an elongated or stretched form on charged optical mapping surfaces; absorbed molecules are then further stabilized by conjugation of a porous polymer overlay (acrylamide gel) [17]. Presented molecules are then barcoded by restriction digestion for later genomic placement or for localization of sequencing products across each long DNA template. This restriction digestion, or barcoding step, is performed only once per sequencing operation.
- Step 2. Template nicking and gapping. The arrayed double-stranded DNA templates are modified for optical sequencing in several steps. First, DNase is added to randomly nick the arrayed DNA molecules; second, a wash step removes nuclease; and finally, addition of T7 exonuclease produces gaps at the nick sites for enabling sequence acquisition using DNA polymerases lacking 5′–3′ exonuclease or strand displacement activities. As optical mapping surfaces are used, enzymatic modifications of bound template molecules are efficient; thus, this nicking step reliably controls the spacing (~1–4 kb) of subsequent SBS reactions, adjusted to be commensurate with the resolution of light microscopy and the degree of DNA template stretching.
- Step 3. Fluorochrome incorporation into nicked/gapped templates. DNA polymerase incorporates fluorochrome-labeled nucleotides within multiple nicked or gapped sites along each template DNA molecule composing an array. Each

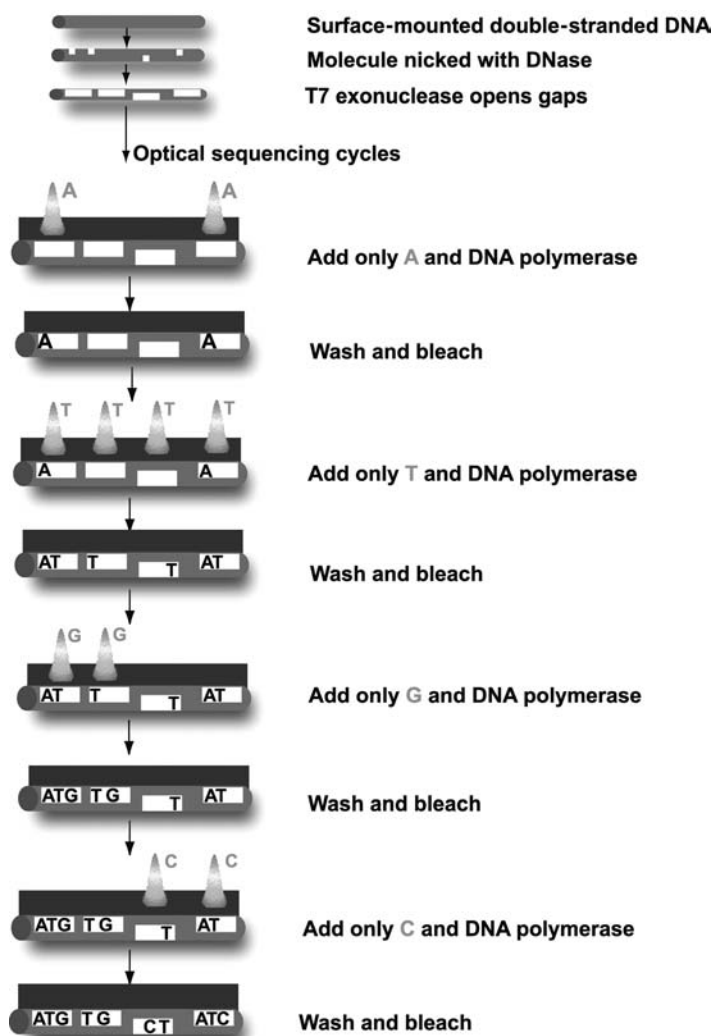


Figure 11.1 Overview of optical sequencing. Step 1: mounting DNA molecules. Single DNA molecules are elongated on optical sequencing surfaces. Step 2: nicking the target. DNase is added to nick surface-elongated target DNA molecules, followed by T7 exonuclease treatment to produce gaps and prepare the template for DNA polymerase nucleotide incorporation. Steps 3–6: optical sequencing cycles –nick translation, gap filling, or strand displacement with labeled

nucleotide. DNA polymerase and fluorochrome-labeled nucleotide (dNTP^f) are added in standard buffers. In each cycle, one type of dNTP^f is incorporated into the gaps, washed, imaged, quantitated, and photobleached. This readies the template for subsequent optical sequencing cycles. The cycle is repeated for each labeled nucleotide until the desired region has been sequenced. (Reprinted from Ref. [10] with permission from Elsevier.)

reaction mix includes only one of the four labeled nucleotides (four mixes: A^f, T^f, G^f, or C^f; “f” indicates fluorochrome labeling), and none, one, or multiple fluorochrome-labeled nucleotide additions could occur at each nick/gap site depending on the template sequence.

- Step 4. Imaging: counting the number of incorporated fluorochromes per nick/gap site. Fluorochrome-labeled nucleotides are imaged as punctates; analysis then counts the number of incorporated fluorochromes for each nick or gap position across all templates. Such counting requires minimal photophysical interactions between incorporated fluorochromes.
- Step 5. Photobleaching. The same laser illumination used for fluorochrome excitation now destroys previously imaged fluorochromes. This step effectively “resets” the previous fluorochrome counting operation, allowing tabulation of labels added in the next cycle.
- Step 6. Repeat Steps 3–6. The details for optical sequencing system provided below have been employed for demonstrating proof-of-principle operation.

11.2.1

Optical Sequencing Microscope and Reaction Chamber Setup

11.2.1.1 Microscope Setup

Figure 11.2 is the schematic drawing of the microscope and reaction chamber setup for optical sequencing. The microscope setup of the optical sequencing system relies on instrumentation designed for simplicity and robust operation. For these reasons, the system was built around a Zeiss Axiovert 135-TV microscope equipped with epifluorescence with laser illumination provided by an argon ion laser (488 nm).

11.2.1.2 Optical Sequencing Reaction Chamber Setup

The reaction chamber setup is basically a metal slide holder for the surface bearing the DNA sample, bound to a fluidic component fabricated from polydimethylsiloxane (PDMS). The reaction chamber has inlets for moving reagents metered by a syringe pump; details are shown in Figure 11.2.

11.2.2

Surface Preparation

Optical sequencing surfaces [10] are similar to optical mapping surfaces created from commercial glass. Rigorous cleaning and derivatization protocols ensure uniform surface modifications that optimize the presentation of genomic DNA molecules and biochemical operations. Details have been provided in optical mapping and sequencing publications [18, 19]. In short, coverslips are cleaned with strong

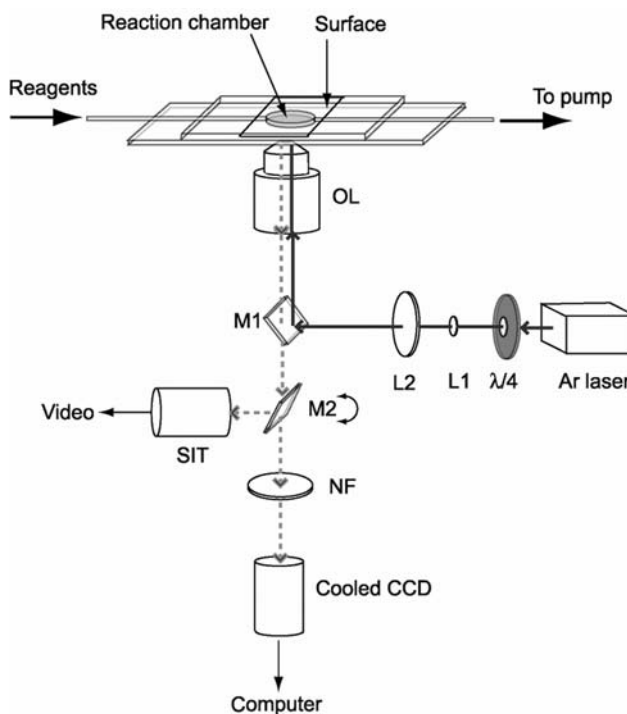


Figure 11.2 Schematic drawing of the optical sequencing microscope setup. A Zeiss Axiocvert 135-TV microscope equipped with epifluorescence was modified as follows: the beam of 488 nm light from an argon ion laser (Ar laser) was circularly polarized by passing it through a quarter wave plate ($\lambda/4$). The beam was then dispersed using lens (L1) and collimated with lens (L2). It was then introduced into a Zeiss Neofluar objective lens (OL), of 1.3 NA, through a dichroic mirror (M1). The surface

bearing the DNA sample was sealed onto a reaction chamber. An inlet for reagents and an outlet are attached to a syringe pump. The fluorescence signal passing through the dichroic mirror (M1) was visualized using a SIT camera or imaged using a CCD camera. The mirror M2 was used to switch between the two cameras. A holographic notch filter (NF) was used to reject the 488-nm excitation light. (Reprinted from Ref. [10] with permission from Elsevier.)

oxidizing agents (piranha; Nano-Strip, Cyantek Corp., Fremont, CA) to remove commercial coatings (preventing sticking) and then boiled in concentrated hydrochloric acid to fully protonate surface silanol groups for subsequent silane coupling steps. Cleaned surfaces are derivatized in an aqueous solution containing trimethyl and vinyl silanes (*N*-trimethylsilylpropyl-*N,N,N*-trimethylammonium chlorides; vinyltrimethoxy-silane; Gelest, Inc., Morrisville, PA). The trimethyl silane contains a positively charged amine group, which provides an anchor for electrostatic interactions between the DNA molecules and the optical mapping surface. The vinyl silane creates covalent cross-links between the acrylamide gel overlay and the optical mapping/sequencing surface.

11.2.3

Genomic DNA Mounting/Overlay

Genomic DNA molecules are elongated and deposited onto optical mapping/sequencing surfaces using a microfluidic PDMS device [15, 20] sealed onto a derivatized surface. After mounting, polyacrylamide is then cured in place, further securing molecules to the surface through fluidic operations.

11.2.4

Nicking Large Double-Stranded Template DNA Molecules**11.2.4.1 Nicking Mounted DNA Template Molecules**

DNase I is used for nicking double-stranded DNA molecules that have been elongated and absorbed on optical sequencing surfaces. The mean number of nicks per template is varied by simple titration of DNase I concentration or by varying the incubation time. The distribution of nick sites is adjusted for spacing them approximately five times the resolution of light microscopy ($\sim 0.2 \mu\text{m}$; 1 mm) or approximately 3 kb of B-form DNA; that is, assuming DNA stretched to 70–90% of the calculated polymer contour length. The activity of DNase I on surface-mounted molecules is efficient and controllable; it has been shown that 0.01–0.0001 units of DNase I and 5 min incubation time are needed for creating usable densities of nick sites for optical sequencing. Thorough washing with TE buffer after DNase I treatment is essential for termination of nicking activity and for minimizing the creation of double-stranded breaks. It is important to note that nicking endonucleases such as Nb.BbvCI and Nt.AlwI (New England Biolabs), which create nicks at specific sequence location, can also be used for nicking the double-stranded DNA templates [16].

11.2.4.2 Gapping Nick Sites

Sequenase v. 2.0, an engineered form of bacteriophage T7 DNA polymerase, is specifically designed for DNA sequencing featuring high processivity and no 3'–5' exonuclease activity. This polymerase also readily incorporates many nucleotide analogues used for sequencing (ddNTPs, thio dNTPs, dITP, etc.) and, most importantly, fluorochrome-labeled nucleotides [11]. As it lacks strand displacement and 5'–3' exonuclease activity, T7 exonuclease is used for producing gaps at nick sites within double-stranded template molecules, thereby punctuating them with single-stranded regions supporting Sequenase action. Accordingly, the extent of gapping by T7 exonuclease must be carefully controlled; too much exonuclease activity will produce an unacceptably high level of double-strand breaks. This issue is resolved by careful titration of nicking activity (as previously described), followed by formation of small-to-moderate gaps. For optical sequencing purpose, only small gaps (20–50 bp) are necessary.

11.2.5

Optical Sequencing Reactions11.2.5.1 **Basic Process**

A unique feature of optical sequencing is that the sequencing cycles do not include steps – whether enzymatic, chemical, or photocleavage – to remove bulky fluorochrome labels after each template-directed incorporation of labeled nucleotides (sequential cycles that incorporate: A^f, T^f, G^f, C^f; repeat). Although obviation of label removal after each nucleotide addition makes sequencing cycles economical, such action requires consecutive additions of bulky fluorochrome-laden nucleotides. Naturally, this action is template directed and ceases after the template requirement at particular given nucleotides was completely satisfied. After the polymerization step has completed, several washes are performed for removal of unincorporated labeled nucleotides, allowing imaging of products and initiation of a new cycle incorporating a new base.

11.2.5.2 **Choices of DNA Polymerases**

Several commercially available DNA polymerases including *Taq*, *Bst*, *Tth*, Klenow (exo-), and Sequenase v. 2.0 have been evaluated for their abilities to consecutively incorporate fluorochrome-labeled nucleotides [11]. The criteria for selection include ability to efficiently incorporate fluorochrome-labeled nucleotides, lack of 3'–5' exonuclease activity (this “proof-reading” activity would remove newly incorporated nucleotides), fidelity of template-directed addition (low misincorporation rates), and good activity using surface-mounted templates. The polymerases tested showed different strengths and weaknesses in terms of fidelity, tolerance of labeled nucleotides, capacity for strand displacement, presence of 5'–3' exonuclease activity, and, most importantly, the ability to consecutively incorporate fluorochrome-labeled nucleotides. As such, this last consideration critically stresses polymerase–fluorochrome interactions, often affected by the length of chemical linkers connecting dyes and nucleotides. R110-dUTP (PE Applied Biosystems, Inc.) was used to establish the baseline nucleotide incorporation conditions as initially assayed by primer extension reactions, since it had been shown to be easily incorporated and its products are readily detected by fluorescence microscopy [17].

11.2.5.3 **Polymerase-Mediated Incorporations of Multiple Fluorochrome-Labeled Nucleotides**

The addition of multiple labeled nucleotide bases raises several inherent experimental concerns; for instance, steric hindrance of bulky fluorochrome moieties may limit addition, and the quantitation of multiple additions could be problematic. These are critical concerns for optical sequencing, because the enzymatic cycles require multiple additions of labeled nucleotides, as they are not serially added, one at a time. In addition, the effective read length depends on the ability to precisely tabulate the number of incorporated fluorochromes through analysis of fluorescence intensity. Here, photophysical interactions between neighboring fluorochromes might

produce nonlinear effects affecting their “counting” by fluorescence intensity measurement of punctates. However, alteration of nucleotide–dye linker length and fluorochrome moieties can deal with this photophysical issue.

We conducted experiments to optimize the high-density, controlled, and sequential incorporation of multiple fluorescently labeled nucleotides into target molecules to explore optimal, single DNA molecule sequencing strategies. Multiple variables were tested including polymerase type, buffer conditions, and fluorochrome chemistries to select optimal conditions for optical sequencing reactions [11].

11.2.5.4 Washes to Remove Unincorporated Labeled Free Nucleotides and Reduce Background

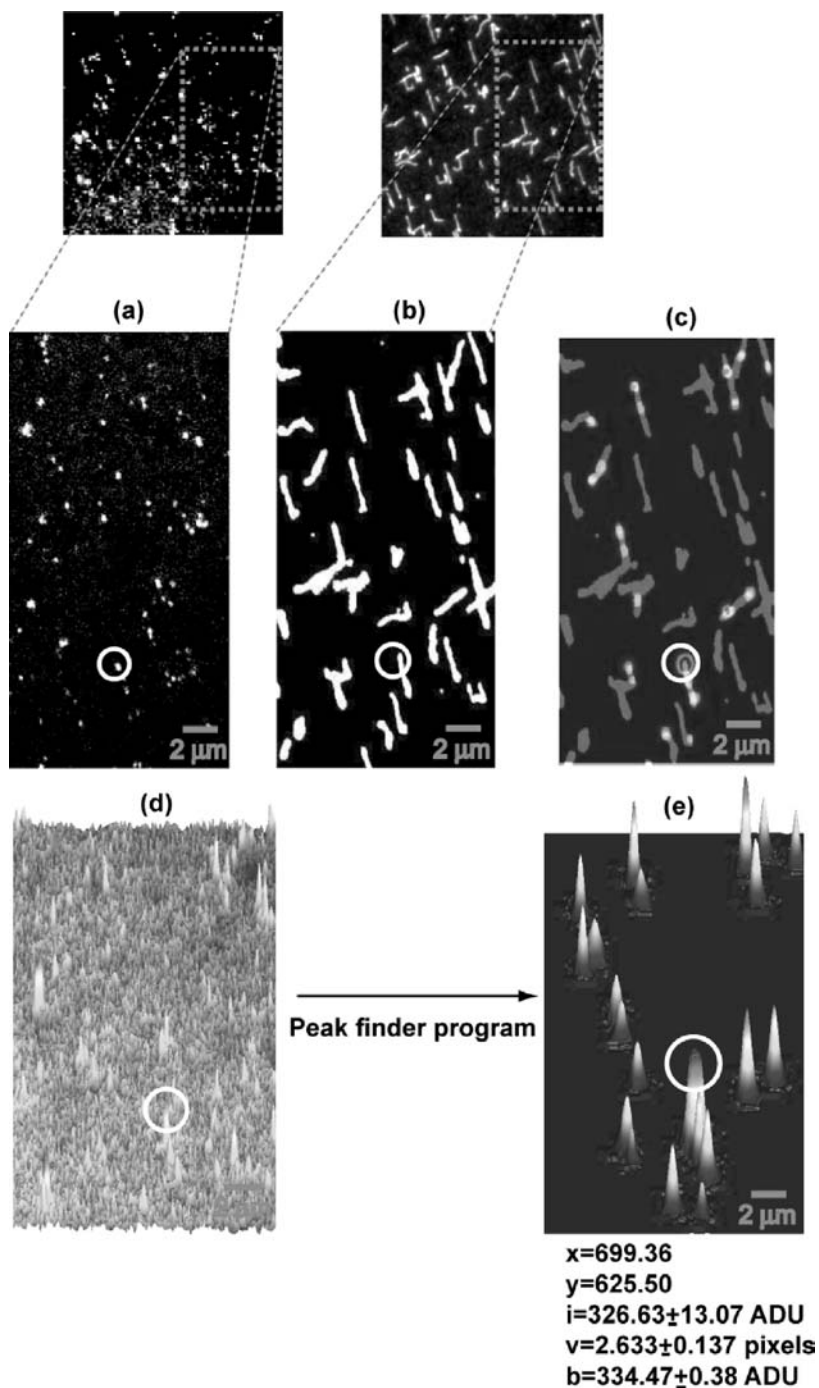
Removal of unincorporated fluorochrome-labeled nucleotides after DNA polymerase incorporation is required as accumulated fluorochromes obscure counting of incorporated labeled nucleotides. The optical sequencing surface is positively charged and can absorb free nucleotides. However, washes with $3 \times \text{SSC}$ (450 mM sodium chloride and 45 mM sodium citrate) can efficiently remove free nucleotides without disturbing the target DNA molecules absorbed on optical sequencing surfaces.

11.2.6

Imaging Fluorescent Nucleotide Additions and Counting Incorporated Fluorochromes

Reliable detection of single fluorochromes is critical for optical sequencing. At present, such measurements are now commonplace using sensitive CCD cameras, “bright” photostable fluorochromes, and microscope imaging techniques such as total internal reflection fluorescence [21, 22]. Results from Schmidt *et al.* showed that single fluorochromes can be localized to within tens of nanometers; furthermore, the Gaussian function is a reasonable approximation of the point spread function (PSF) [23]. Guided by these early findings, we incorporated noise suppression into our punctate analysis approach by using an empirically determined point spread function (using 20 nm fluorescent latex beads) that was automatically fitted to profiles generated across all detectable fluorescence punctates developed through cycles of optical sequencing. Those fluorescence intensity profiles that did not fit the imaging system PSF well were classified as “noise” and rejected [10]. Further filtering of noise was developed through a scheme of “addition histories,” which discriminated “false” punctates based on expectations of fluorescence intensities as sequencing cycles accumulate (Figure 11.5; further discussed in Section 11.2.8).

To demonstrate single-fluorochrome detection and Gaussian fitting of associated punctate products, a 10 kb PCR amplicon (λ –bacteriophage template), bearing a known distribution of fluorochromes, was used [10]. Figure 11.3a shows a subsection of a raw image obtained of the R110-labeled amplicons. Following the imaging of R110 punctates, molecules were stained with YOYO-1 and imaged again



(Figure 11.3b); superimposition of these two image planes allowed the identification of punctate-associated DNA molecule “backbones” (Figure 11.3c). Accordingly, the analysis of fluorochrome signals (punctates) employed the Peakfinder program for fitting data to a two-dimensional Gaussian model. Figure 11.3d depicts a three-dimensional representation of the data of panel (a), and Figure 11.3e shows the results of Peakfinder analysis on the signals in the 3D R110 fluorochrome image. The signals that fit the model were first filtered for those that matched the PSF (the expected signal for diffraction-limited fluorescent punctates) and second for colocalization with a DNA backbone (signals must emanate from a template). The Gaussian parameters used to characterize the signals included x —the x -coordinate of the signal; y —the y -coordinate; i —the total counts of each peak; v —the variance; and b —the reduced χ^2 value of the fit and background. The Gaussian fit parameters are listed for a representative peak that satisfied the filters (Figure 11.3e). (The peaks selected for further analysis were those that fit the Gaussian model and had a full-width-at-half-maximum (FWHM) of 269.96 ± 144.8 (two standard deviations from the mean PSF value).)

The criteria that served as evidence that a signal was from a single R110 incorporated into the DNA molecule included the single fluorochrome signal needed to match the PSF of a point source, the single fluorochrome peak needed to colocalize with the YOYO-1 stained DNA backbone, and the single fluorochrome needed to photobleach in one step. Over 7000 punctates were measured whose fluorescent signals matched the PSF of a point source; approximately 70% of these (4925) also colocalized with the DNA backbone. These were used to characterize single fluorochromes (Figure 11.4). The model fitting provided the intensity of the single-fluorochrome peaks (i) with the background value (b). To correct for the nonuniform illumination in the image, the i/b ratio was used as an arbitrary value to normalize the peak intensities.

The primary distribution seen in Figure 11.4a (indicated by arrow 1A and representing the distribution contributed by PCR products with one fluorochrome i/b value) has a mean i/b value of 0.83 ± 0.24 . A subpopulation of the PCR products will have two fluorochromes incorporated; 1.5% of the total is expected to have two. If they are within the diffraction limit, or if DNA molecules with multiple

Figure 11.3 Imaging and characterization of single fluorochromes. Images of R110-labeled lambda PCR products were acquired using 300 mW/cm² (488 nm, Ar ion laser) and 1200 ms illumination times. (a) Subsection of a raw image obtained from imaging the R110-labeled DNAs. (b) DNA backbones were imaged following YOYO-1 staining; an identical subsection of the same field as in (a) is shown. The fluorochrome signals were model fitted to a two-dimensional Gaussian function, using the Peakfinder program. (d) Three-dimensional representation

of a. The data shown in (e) are the result of analysis of the data of (d) with the Peakfinder program. The signals that fit the model were then filtered for those that matched the PSF. A mask of (b) was made and overlaid onto the image in (a) and the signals were further filtered on the basis of those that colocalized with a DNA backbone (c). The encircled peak, which passed through these filters, yielded the indicated Gaussian fit parameters. (Reprinted from Ref. [10] with permission from Elsevier.)

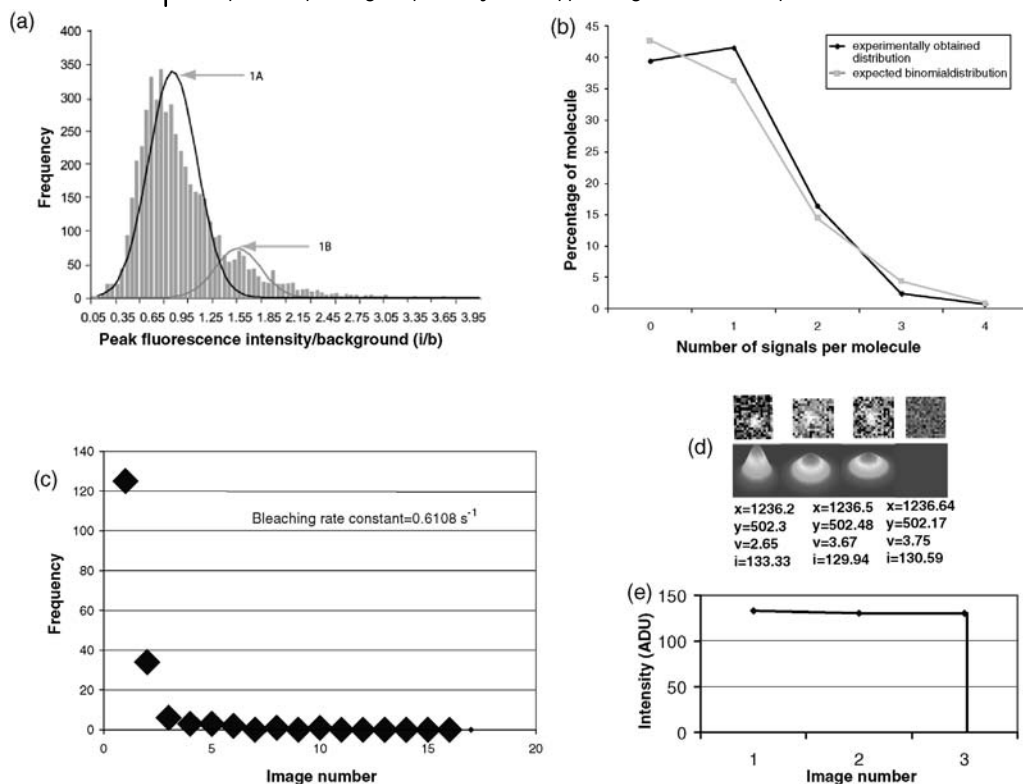


Figure 11.4 (a) Intensity distribution of signals obtained from R110-labeled PCR products. Histogram of peak (intensity/background) values of 4925 signals from R110-labeled PCR products, at 300 mW/cm^2 of 488-nm laser light, and an exposure time of 1200 ms. The arrows 1A and 1B indicate distributions contributed by PCR products with one and two fluorochrome i/b values, respectively. (b) Comparison of the expected binomial distribution versus the experimentally obtained data of the number of signals per DNA molecule. Distribution of PCR-amplified DNA molecules that colocalize with zero, one, two, three, and four single-fluorochrome signals. For the experimentally obtained distribution series, 1616 DNA molecules were analyzed. The expected binomial distribution series is the expected distribution of

the corresponding number of dyes. (c–e) Single fluorochromes are bleached in a single step. Multiple images of fluorochrome-labeled PCR products were acquired at 400 ms intervals. Analysis of these images (c) showed a simple exponential decay and a fit produced an apparent bleaching rate constant (0.6108 s^{-1}). The images in the top panel of (d) are the raw images of a single fluorochrome. The images that are below it show the corresponding fit, while the fourth frame showed no detectable signal. The Gaussian fit parameters of the peaks in (d) are indicated below the images. Panel (e) contains a plot of the intensity of the peaks plotted as a function of image number, with the measured intensity dropping below the measurement after frame 3. (Reprinted from Ref. [10] with permission from Elsevier.)

fluorochromes are not completely elongated, they will appear as one fluorochrome. This could also occur if more than one labeled DNA is colocalized. The second distribution observed has a mean i/b value of 1.50 ± 0.231 (for a total of 738 signals; Figure 11.4a, noted by the arrow 1B). The i/b value of this distribution was

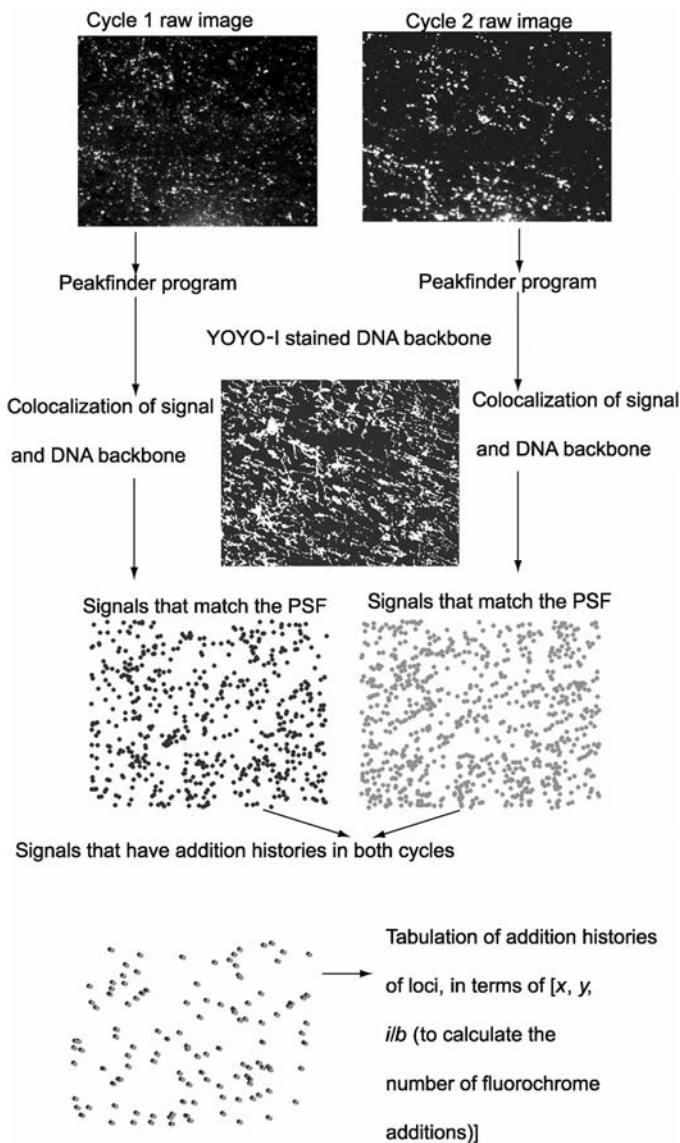


Figure 11.5 Analysis of the addition histories obtained from an optical sequencing cycle. Flowchart of the analysis of the optical sequencing cycle data obtained an ensemble of molecules in an image following R110-dUTP addition (cycle 1) and R110-dCTP addition (cycle 2). The addition histories of each locus were characterized by the x and y coordinates of the

R110 addition signals from the two cycles that matched the PSF, colocalized with each other and with the DNA backbone, and the intensity/background (i/b) value. Here, x, y stand for the x - and y - coordinates of the Gaussian fit, i is the intensity, and b is the background value obtained from the Gaussian fit. (Reprinted from Ref. [10] with permission from Elsevier.)

approximately twice that expected from a single fluorochrome, which suggested that the signal was due to two fluorochromes within the diffraction limit. The reason that this distribution had more observed signals ($\sim 20\%$ of the total) than the expected 1.5% might be the uneven stretching of lambda molecules on the surface. This could result in the apparent increase in the number of molecules with two fluorochromes within the diffraction limit. The signal-to-noise (S/N) ratio was determined to be 4.20 for a single fluorochrome. On the basis of the analysis provided by Schmidt *et al.* [23], the stoichiometric resolution, or the potential to count the number of fluorochromes, was determined as follows. For n colocalized fluorochromes of intensity to background ratio $(i/b)_n$ and standard deviation σ_n , the expected fluorescence with regard to the intensity-to-background ratio $(i/b)_1$ and standard deviation σ_1 of one fluorochrome can be written as $(i/b)_n \pm \sigma_n = n(i/b)_1 \pm \sqrt{n}\sigma_1$.

If $(i/b)_1$ is known, then the number n can be determined as long as σ_1 is smaller than $n(i/b)_1$ or n is smaller than $[(i/b)_1/\sigma_1]^2$. Therefore, $[(i/b)_1/\sigma_1]^2$ sets the limit for the number of fluorophores that can be counted. From Figure 11.4a, we know the values of $(i/b)_1$ and σ_1 , 0.83 and 0.24, respectively. The stoichiometry for the quantitation of colocalized dyes was determined to be approximately 12, calculated on the basis of the assumption that fluorochromes do not interact with each other. In reality, though, when two fluorochromes are too close to each other, they could interact with each other, which would result in a nonadditive nature of the fluorochrome intensities.

Further analysis was done to show that colocalization of the fluorochrome signals and the DNA backbones is not random but is evidence of a biochemical event [10]. The total image area occupied by the single DNA molecules was sampled for the occurrence of signals, and the fraction of the total image area occupied by the DNA molecules (defined as P1) and that of the R110 signals that colocalize with DNA molecules (defined as P2) were calculated. If the distribution of signals is random, P2 should equal to P1. However, for the R110-labeled PCR products, P1 was calculated to be 0.14 and P2 to be 0.63. The fact that P2 is over four times larger than P1 strongly supports the view that the signals obtained were not due to random association. The expected binomial distribution and the experimentally obtained distribution are seen to be in close agreement in the graph of Figure 11.4b. These data were obtained from analysis of 4914 single-fluorochrome signals (from 1616 molecules). The experimentally obtained distribution showed 41.5% of molecules with one signal; the expected binomial distribution was 36.3%. This small difference could be due to imperfect stretching of DNA backbones, which would result in an increase in the colocalization of two signals.

Additional analysis to assess whether the measured signals came from single fluorochromes was conducted in a bleaching series with the R110-labeled lambda PCR products. The bleaching series used an illumination intensity of approximately 300 mW/cm^2 and an illumination time of 400 ms per image [10] (Figure 11.4c–e). The average S/N ratio was found to be 9.22, determined from measurements of single fluorochromes taken after the 400 ms exposure; this was

about half that at the exposure time of 1.2 s. The peak intensity of the single fluorochromes (as analyzed using the Peakfinder program) was found to disappear within a single step, as demonstrated in Figure 11.4d and e. These findings strongly suggest that single fluorochromes were the source of the observed fluorescence signals (Figure 11.4d). The bleaching rate constant was determined to be 0.6108 s^{-1} (Figure 11.4c) from data collected from a population of molecules; half of these would be bleached within 1.1345 s. These experiments confirm that single-fluorochrome detection is robustly done by using conventional fluorescence microscopy (Figure 11.2).

11.2.7

Photobleaching

Removing fluorochrome signals after their imaging and quantitation is an essential part of optical sequencing. The main reason for photobleaching after addition and imaging is to eliminate any carryover of fluorescence signals between cycles. One advantage of using this process is that it is nonenzymatic, is rapid, works in virtually any buffer, and does not require addition or subtraction of reagents. As previously discussed, Figure 11.4c–e shows that single-step photodestruction of a single fluorochrome rapidly occurs in approximately 1.2 s [10], which is approximately the time needed to take three images.

Although the photobleaching step is simple and effective, collateral damage to the DNA template may attenuate further nucleotide incorporation; the next section describes experimental findings showing that such template damage is not an issue.

11.2.8

Demonstration of Optical Sequencing Cycles

An optical sequencing cycle consists of incorporating, imaging, and reading of the dNTP_fs and a photobleaching step to “reset” the fluorochrome counting operation. Other approaches for single-molecule sequencing have relied on schemes to remove labels after addition to facilitate the counting of fluorochromes. As potential photodamage of the template during photobleaching would attenuate the ability of polymerase to add subsequent nucleotides, particularly when a fluorochrome is present on the nascent strand, this concern was addressed by a series of primer extension reactions using lambda DNA as template bearing a known distribution of fluorochromes. Here, templates were shown to support primer extension after photobleaching steps [11] using the instrumentation shown in Figure 11.2 and gel electrophoresis, which showed no appreciable diminution of extension products [11].

The optical sequencing schema, shown in Figure 11.1, was realized through experiments that combined all of the system components described in this review. First, experiments, shown in Figure 11.5, were performed demonstrating the

utility of the addition histories – a scheme that tracks and filters weak fluorescent signals on the basis of locus and fit to the PSF. Accordingly, Figure 11.6 shows a series of images and analyses of two optical sequencing cycles performed on lambda DNA templates after DNase nicking and gapping with T7 exonuclease. In brief, the punctate signals that colocalized with the DNA backbone in each cycle were identified. These filtered signals were then correlated between the two cycles for identifying those gaps supporting successive incorporation of R110-dUTP, followed by R110-dCTP.

11.3

Future of Optical Sequencing

Optical sequencing was conceived for the acquisition of large data sets comprising short strings of sequence information, derived from very long DNA templates. Although published work on optical sequencing now pales in comparison to current next-generation sequencing platforms, acquisition of short strings of sequences across long double-stranded DNA templates remains a powerful advantage, as independently barcoded molecules support genomic placement, irrespective of associated sequence information, while uniquely revealing structural variation and somatic mutation. In part, these advantages obviate many of the issues plaguing human genome analysis by sequencing platforms offering high-throughput operation but modest read lengths. Advances in detection, engineered polymerase, and dye chemistries have converged so that these developments will enable new single-molecule sequencing platforms to meet the needs of population-based genetics and diagnostics.

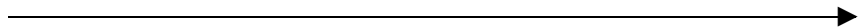
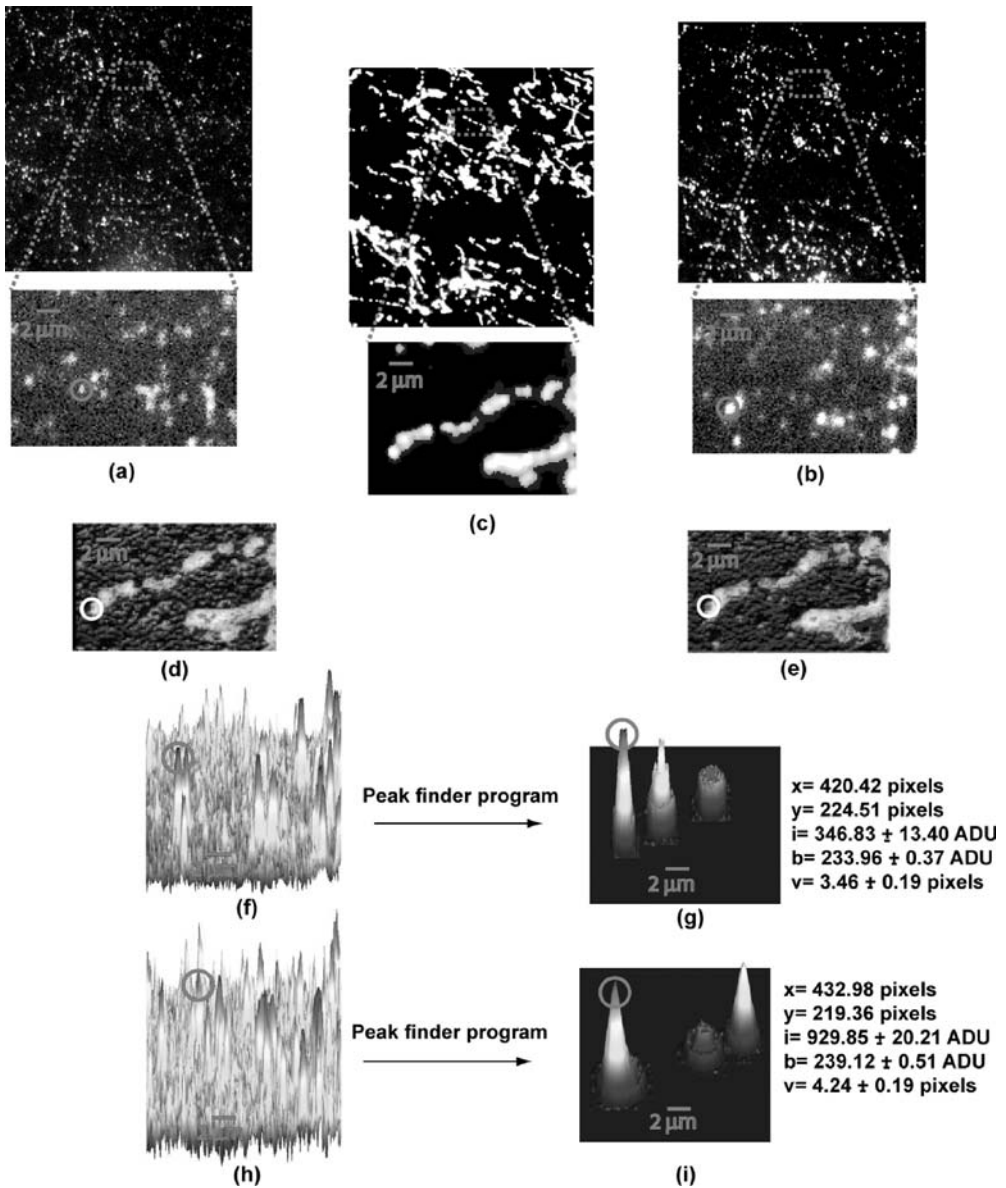


Figure 11.6 Demonstration of an optical sequencing cycle. The optical sequencing cycle was carried out on the lambda DNA templates that were elongated and absorbed onto derivatized surfaces, nicked with DNase I, and gapped with T7 exonuclease. Sequenase v. 2.0 was challenged with dATP, dCTP, and dCTP, to reset all the gaps for R110-dUTP addition, and surfaces were washed. (a) Subsection of the image of R110-dUTP additions obtained from challenging Sequenase v. 2.0 with only R110-dUTP (upper left panel). After washing and photobleaching, Sequenase v. 2.0 was challenged with dATP, dGTP, and dTTP, to reset the gaps for addition of R110-dCTP, and the surface washed. Following this, the second cycle was initiated by addition of DNA polymerase and only R110-dCTP. (b) Image from the same location as subsection (a), after addition of R110-dCTP in the second cycle; (c) image of the YOYO-1 stained DNA from the same subsection. (d and e) Overlays indicating the colocalization of the R110 signals and the DNA backbone. (f and h) The data from a and b are represented three dimensionally. (g and i) Result of the Peakfinder program on f and h, respectively. (g) The encircled peak represents the addition of R110-dUTPs in the first cycle. (i) The encircled peak shows the addition of R110-dCTP in the same locus in the second cycle. The signals yielded the indicated Gaussian parameters after Peakfinder analysis.



References

- 1 Schwartz, D.C. (2004) The new biology. in *The Markey Scholars Conference* (ed. G.R. Reinhardt), National Academies Press, Puerto Rico, pp. 73–79.
- 2 Steemers, F.J. and Gunderson, K.L. (2007) Whole genome genotyping technologies on the BeadArray platform. *Biotechnology Journal*, 2, 41–49.

- 3 Shendure, J., Mitra, R.D., Varma, C. and Church, G.M. (2004) Advanced sequencing technologies: methods and goals. *Nature Reviews. Genetics*, **5**, 335–344.
- 4 Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K. *et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
- 5 Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, **242**, 84–89.
- 6 Ronaghi, M., Uhlen, M. and Nyren, P. (1998) A sequencing method based on real-time pyrophosphate. *Science*, **281**, 363, 365.
- 7 Leamon, J.H., Lee, W.L., Tartaro, K.R., Lanza, J.R., Sarkis, G.J., deWinter, A.D., Berka, J., Weiner, M. *et al.* (2003) A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, **24**, 3769–3777.
- 8 Braslavsky, I., Hebert, B., Kartalov, E. and Quake, S.R. (2003) Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 3960–3964.
- 9 Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G. and Webb, W.W. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, **299**, 682–686.
- 10 Ramanathan, A., Huff, E.J., Lamers, C.C., Potamouis, K.D., Forrest, D.K. and Schwartz, D.C. (2004) An integrative approach for the optical sequencing of single DNA molecules. *Analytical Biochemistry*, **330**, 227–241.
- 11 Ramanathan, A., Pape, L. and Schwartz, D.C. (2005) High-density polymerase-mediated incorporation of fluorochrome-labeled nucleotides. *Analytical Biochemistry*, **337**, 1–11.
- 12 Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, **18**, 630–634.
- 13 Valouev, A., Li, L., Liu, Y.C., Schwartz, D.C., Yang, Y., Zhang, Y. and Waterman, M.S. (2006) Alignment of optical maps. *Journal of Computational Biology*, **13**, 442–462.
- 14 Valouev, A., Schwartz, D.C., Zhou, S. and Waterman, M.S. (2006) An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 15770–15775.
- 15 Zhou, S., Herschleb, J. and Schwartz, D.C. (2007) A single molecule system for whole genome analysis, in *New High Throughput Technologies for DNA Sequencing and Genomics* (ed. K.R. Mitchelson), Elsevier, Amsterdam, pp. 269–304.
- 16 Jo, K., Dhingra, D.M., Odijk, T., de Pablo, J.J., Graham, M.D., Runnheim, R., Forrest, D. and Schwartz, D.C. (2007) A single-molecule barcoding system using nanoslits for DNA analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2673–2678.
- 17 Hu, X., Aston, C. and Schwartz, D.C. (1999) Optical mapping of DNA polymerase I action and products. *Biochemical and Biophysical Research Communications*, **254**, 466–473.
- 18 Lim, A., Dimalanta, E.T., Potamouis, K.D., Yen, G., Apodoca, J., Tao, C., Lin, J., Qi, R. *et al.* (2001) Shotgun optical maps of the whole *Escherichia coli* O157:H7 genome. *Genome Research*, **11**, 1584–1593.
- 19 Zhou, S., Deng, W., Anantharaman, T.S., Lim, A., Dimalanta, E.T., Wang, J., Wu, T., Chunhong, T. *et al.* (2002) A whole-genome shotgun optical map of *Yersinia pestis* strain KIM. *Applied and Environmental Microbiology*, **68**, 6321–6331.
- 20 Dimalanta, E.T., Lim, A., Runnheim, R., Lamers, C., Churas, C., Forrest, D.K.,

- de Pablo, J.J., Graham, M.D. *et al.* (2004) A microfluidic system for large DNA molecule arrays. *Analytical Chemistry*, **76**, 5293–5301.
- 21 Nie, S., Chiu, D.T. and Zare, R.N. (1994) Probing individual molecules with confocal fluorescence microscopy. *Science*, **266**, 1018–1021.
- 22 Funatsu, T., Harada, Y., Tokunaga, M., Saito, K. and Yanagida, T. (1995) Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution. *Nature*, **374**, 555–559.
- 23 Schmidt, T., Schutz, G.J., Baumgartner, W., Gruber, H.J. and Schindler, H. (1996) Imaging of single molecule diffusion. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 2926–2929.

12

Microchip-Based Sanger Sequencing of DNA

Ryan E. Forster, Christopher P. Fredlake, and Annelise E. Barron

For decades, the method of choice for the determination of DNA sequences has been the Sanger reaction, followed by size-based electrophoretic separation of single-stranded DNA ladders. The development of capillary array electrophoresis (CAE) [1, 2] to replace the more traditional slab gels [3] certainly led to dramatically increased DNA sequencing throughput, but sequencing human genomes by this technology was still far too expensive (\$20 million per individual). Further advances in sequencing technology are needed to drive down the cost per sequenced base, allow sequencing data to be made available to biological and medical researchers more quickly, and facilitate both research and genomic medicine (“personalized medicine”). While many new technologies for the determination of DNA sequences are under development, at this time, Sanger sequencing is the only technology that can provide truly long “reads” (i.e., the highly accurate sequence of more than 600 contiguous DNA bases). As such, this technology is not going to go away any time soon. Current development of more advanced Sanger-based electrophoretic sequencers involve the miniaturization of the process onto a microfluidic chip platform [4–8]. Microfluidics has the potential to greatly reduce costs in each step of the sequencing process from sample preparation to analysis. These devices can produce more defined, narrower sample injection zones, which increases DNA separation efficiency and thus decreases the total distance and time required to obtain high-resolution data [9]. Microfluidic devices can also be designed and fabricated so that every step from sample preparation to separation can be performed on one integrated device, and so that up to 96 samples can be processed simultaneously [10, 11].

Miniaturizing and combining multiple processes onto a microfluidic device offers obvious advantages; however, the engineering of multiple chemical processing steps onto a small device has created new challenges and problems not encountered in comparative macroscale benchtop systems [12–17]. Here, we will discuss advances made recently in the development of polymeric materials for high-resolution sequencing separations on chips, as well as some notable examples of novel microfluidic systems for Sanger-based sequencing technology.

12.1

Integrated Microfluidic Devices for Genomic Analysis

The development of a single microfluidic device capable of pre-PCR DNA purification, amplification *via* thermal cycling, post-Sanger reaction purification, electrophoretic separation, and finally DNA detection will play a significant role in the pursuit of rapid, inexpensive genomic sequence determination. Combining these steps onto a single microfluidic platform promises to greatly reduce the amount of expensive reagents and materials needed for sequencing and shorten the overall analysis time. The Mathies group at the University of California at Berkeley and the Landers group at the University of Virginia have focused on developing prototypes of microfluidic systems capable of achieving this goal through a combination of glass fabrication techniques, microfluidic valving in PDMS, hydrodynamic pumping, and electrophoresis. Interestingly, these two groups have approached this problem using entirely different techniques.

The integrated devices developed by the Mathies group have been tested with DNA samples that have already been purified from their raw state (i.e., whole blood, serum, etc.). Using resistive heaters integrated into the chip device itself, the DNA sequencing ladder is synthesized *via* the Sanger cycle sequencing reaction with a predetermined set of reagents [10, 18]. To achieve high-resolution separations, DNA must be separated from the extraneous Sanger reaction components before it can be analyzed. The Mathies lab device uses acrylamide-based copolymers with single-stranded oligonucleotides randomly attached to the polymer backbone. By electrophoresing the DNA sample through the polymer, single-stranded DNA with a complementary sequence to the immobilized “capture” oligonucleotides selectively hybridizes while the unwanted molecules (salt, dNTPs, and ddNTPs) are electrophoresed away [11, 19]. By raising the temperature of the device above the melting point of the captured DNA, the desired Sanger fragments can then be released into another channel for electrophoretic separation and detection *via* laser-induced fluorescence (LIF) with four emission channels (“colors”) to detect each DNA base, as required for DNA sequencing. A representative device developed by the Mathies group can be seen in Figure 12.1 [10]. In recent advancements, this system has been modified to allow “in-line injection” of the sample, where the DNA is captured in the same channel in which the electrophoretic DNA separation will occur [19]. This eliminates the need for excess DNA sample, much of which is often wasted during the standard cross-injection in microchip systems, and hence is a step toward significantly reducing sample and reagent requirements by exploiting microfluidics. The Mathies group has also developed microchannel devices with up to 96 sequencing lanes running in parallel, which in principle are capable of sequencing over 100 000 bases per hour when full automation can be achieved [20]. The combination of these two technologies displays the potential of microfluidics to prepare and analyze numerous samples in extremely short periods of time compared to conventional CAE systems. Seamless, robust integration and automation of these processes is the next great challenge that will be faced in developing the chip system invented in the

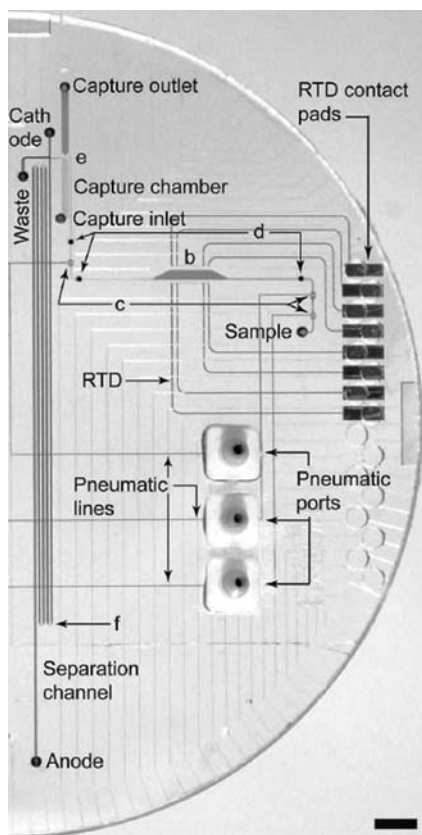


Figure 12.1 Representative photograph of the Mathies integrated DNA sequencing device. The individual parts of the device are labeled in the figure: (b) thermal cycling reactor; (c) microvalves; (d) *via* holes; (e) capture-inject region; (f) tapered turns for separation channel. All features are etched to a 30 μm depth. The scale bar is 5 mm. Reproduced with permission from Ref. [10].

Mathies lab, and this presently is being undertaken by the company Microchip Biotechnologies, Inc. (Dublin, CA).

An entirely different method for analyzing genomic material has been developed by the Landers group. This single-channel device, designed primarily for medical or forensic applications, has the ability to start with a crude biological sample, such as whole blood, and, in less than 30 min, to present a user with an electropherogram representing the size of a specific target DNA fragment [21]. The cells in the crude sample are lysed, releasing the genomic material, and the mixture is purified by hydrodynamically pumping the lysis solution through a silica sol-gel monolith or a photopolymerized monolith synthesized from 3-(trimethoxysilyl) propyl methacrylate [22, 23]. Through a process known as solid-phase extraction (SPE), the DNA

physically adsorbs to the monolith surfaces while most of the cellular debris flows through the bed [24, 25]. After all impurities and PCR-inhibiting molecules are flushed away, the DNA is eluted from the column by pumping a different buffer, such as 10 mM Tris, 1 mM EDTA (TE), through the system. The solution containing the purified DNA is then pumped into a PCR chamber and the target region of the DNA sample is amplified by rapid thermal cycling of the chamber. Instead of using resistive heating, a technique known as infrared-mediated noncontact heating is employed [26–28]. This method was pioneered by the Landers group and provides extremely rapid thermal cycling for DNA amplification; it also eliminates the need for microfabricated heating systems, which can greatly increase the cost of a device. Once the DNA is amplified, the solution is pumped into an electrophoretic separation channel that ends in a single-color LIF detection system for forensic analysis and pathogen detection. This system has not yet been used for DNA sequencing; however, with a few changes to its setup and the reagents used for amplification, as well as the detector, it would have the ability to sequence specific targets from crude biological samples, advancing the development of point-of-care medical or biological detection systems.

12.2

Improved Polymer Networks for Sanger Sequencing on Microfluidic Devices

12.2.1

Poly(*N,N*-dimethylacrylamide) Networks for DNA Sequencing

Developing lab-on-a-chip systems has been a major focus of the DNA sequencing community that is developing Sanger technology, and some very significant advances have been achieved; however, in general, much less attention has been paid to DNA separation networks and polymeric channel coatings utilized in these devices, compared to the development of the devices themselves. Typically, the same materials that were successfully utilized in CAE systems have been used in microfluidic devices; however, just as cross-linked polyacrylamide or agarose networks, which performed extremely well in slab gels, did not transfer well to capillary systems, CAE-specific polymer solutions also need to be reengineered for the new microchip platforms.

To achieve DNA sequencing read lengths of 600–700 bases, which are necessary for current DNA sequence alignment algorithms to process repeat-rich genomes [29, 30], highly entangled solutions of hydrophilic, high-molar mass polymers are needed [31, 32]. Highly hydrophilic polymer coatings for internal microfluidic channel surfaces are also necessary to reduce electroosmotic flow and bioanalyte adsorption, which otherwise greatly reduce the read length and resolution obtained in these devices [33]. Poly(*N,N*-dimethylacrylamide) (pDMA) used in conjunction with poly(*N*-hydroxyethylacrylamide) (pHEA) as a separation matrix and wall coating, respectively, has been reported to provide chip-based read lengths in excess of 600 bases in only 6.5 min in a 7.5-cm long glass channel [34]. These results are 2–3 times

faster than comparable read lengths obtained in other microfluidic chips [6, 7, 10, 35] and 10–20 times faster than a typical CAE system, which requires 1–2 h [31, 36]. A mixture of both high- and low-molecular weight polymers was used to achieve optimal separation of both small and large DNA fragments, as has been demonstrated by Karger and coworkers [36]. The combination of differently sized polymers allows for an increase in sequencing read length because higher total polymer concentration, the most important factor in separating small DNA fragments, results in a smaller average mesh size, while an increase in the polymer entanglement strength, which is achieved with the higher molecular weight polymers, favors optimal separations of larger DNA fragments. Mixed molar mass pDMA matrices provide average read lengths that are 10% longer than matrices formulated with a single average molar mass. Interestingly, a commercially available linear polyacrylamide (LPA) solution from Amersham was tested under the same conditions as the mixed molar mass pDMA and produced less than 300 bases of good-quality data. This is a surprising result since this commercial LPA matrix can often deliver read lengths in excess of 700 bases in a CAE instrument. However, the pDMA matrix shows a significant reduction in band broadening, that is, the dispersion DNA fragment peaks undergo during electrophoresis, compared to the commercial LPA matrix, as shown in Figure 12.2. This reduction in peak width contributes to the ability of these matrices to achieve the more efficient separations and increased read lengths within the shorter channel lengths typically used in microfluidic chips.

The increase in sequencing performance in pDMA is attributed to a hybrid separation mechanism that has been observed *via* single-molecule fluorescent DNA imaging [34, 37, 38]. Theory on DNA electrophoresis through gels and entangled

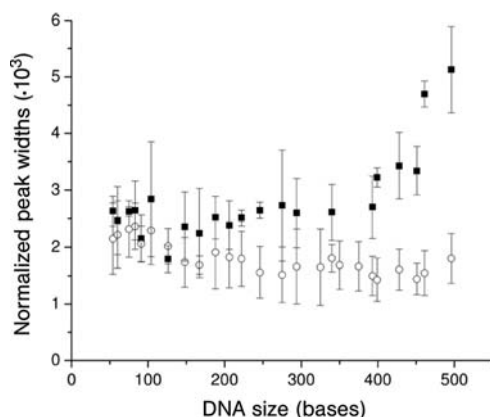


Figure 12.2 Analysis of normalized peak widths of T-terminated fragments of an M13 sequencing sample with DNA size. LongRead LPA sequencing matrix from Amersham (closed squares) is compared with a 4% mixed molar mass pDMA solution (open circles). Peak widths were measured in units of time and normalized by the elution time of the fragment from the microchannel. Reproduced with permission from Ref. [34].

polymer solutions postulates that DNA moves through entangled polymer networks either in an equilibrium coiled conformation (so-called Ogston sieving [39, 40]) or by unwinding and “snaking” through the mesh by a mechanism related to polymer reptation [41, 42]. Dilute polymer solutions, however, can separate DNA by a mechanism known as transient entanglement coupling (TEC) in which DNA entangles with loose polymer chains in solution and transiently drags them through solution in a U-shaped conformation [43–45]. The pDMA matrix discussed above is thought to allow DNA to move through the matrix by a hybrid mechanism using a combination of reptation and TEC (for DNA molecules too large to move through the polymer mesh without unwinding). The pDMA chains form an entangled network in solution that promotes reptation, but the chain entanglements are weak enough to allow the DNA to pull polymer chains from the network, resulting in local network disruption, so that they move through the matrix in a U-shaped conformation similar to what is observed in the TEC mechanism. Figure 12.3 shows time evolution images captured by single-molecule fluorescent imaging of λ -DNA labeled with YOYO-1, electrophoresing through a 3% (w/w) pDMA matrix at room temperature. One of the

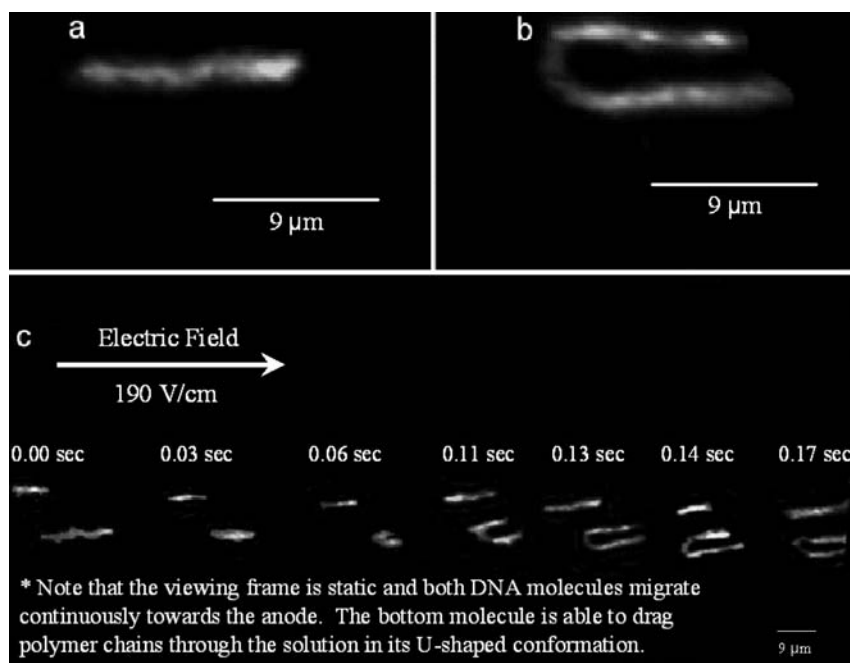


Figure 12.3 Representative images captured by single-molecule fluorescent imaging of λ -DNA labeled with YOYO-1. (a) λ -DNA reptating through an entangled pDMA network. (b) Single λ -DNA molecule that has entangled with the pDMA polymer matrix and extended to form a U-shaped conformation. The molecule is dragging the disentangled pDMA strands through the solution. (c) A series of time-lapse video frames that show two DNA molecules moving through the pDMA network (same molecules in (a) and (b)). Reproduced with permission from Ref. [34].

molecules is reptating through the entangled pDMA network, while a second molecule has entangled with the pDMA matrix and is dragging the disentangled polymer strands through the solution, in a conformation similar to the TEC separation mechanism. While these are not DNA sequencing fragments, they do show that the two mechanisms can coexist in one matrix, under relevant field strengths. The reduction in band broadening and increased read lengths achieved with the mixed molar mass pDMA solutions on glass microfluidic chips results in a polymer separation matrix that is capable of providing long DNA sequencing read lengths, very rapidly, in short separation distances.

12.2.2

Hydrophobically Modified Polyacrylamides for DNA Sequencing

Another advancement in the development of DNA sequencing matrices for microfluidic chips has been made by modifying LPA with hydrophobic *N,N*-dialkylacrylamides to create a hydrophobically modified block copolymer, which is used in conjunction with a pHEA channel coating [46]. To synthesize a reproducible block structure, free-radical micellar polymerization is used, in which a surfactant such as sodium dodecyl sulfate (SDS) is added to the polymerization reaction above its critical micellar concentration [47–50]. The SDS molecules form micelles with hydrophobic cores where the hydrophobic monomer can be incorporated and then effectively integrated into the polymer backbone during the polymerization process [51, 52]. Even with the incorporation of only ~0.1 mol% of *N,N*-dihexylacrylamide monomer into the copolymer, these matrices provide up to a 10% increase in average DNA sequencing read length over LPA homopolymers of matched molar mass. This increase in read length has been attributed to the intermolecular and intramolecular physical cross-linking that occurs between the hydrophobic blocks on the copolymer chains. This effect likely simulates either a larger molecular weight homopolymer or a mixed molar mass solution resulting in the longer average sequencing read lengths. On a 7.5-cm glass offset T microfluidic chip, an average of 554–583 bases/run have been sequenced in 9.5–11.5 min using 4% (w/w) solutions of the copolymer, depending on the molecular weight of the copolymer (ranging from ~1.4 to 7.3 MDa), with the longest sequencing read in excess of 600 bases (98.5% accuracy).

The ability to rapidly load and unload a DNA separation matrix in a microfluidic chip is necessary to analyze multiple DNA samples quickly. Hydrophobic block copolymers typically have much higher viscosities than their homopolymer counterparts [53]; however, due to the extremely small amount of hydrophobe needed to achieve an increase in read length, these specially designed copolymers have very similar viscosities and channel loading times to their homopolymer counterparts. For example, the average loading time for a 4% (w/w), 1.4 MDa (M_w) LPA solution using 200 psi of pressure is ~1 min, and for a matched molar mass LPA-co-DHA copolymer solution, the time is increased by only ~15 s. This is attributed to both the small amount of hydrophobe present in the copolymer and the large extent of shear thinning observed with these copolymers. Shear thinning is especially pronounced

in these copolymer networks because the physical cross-links can break and re-form when the solution is placed under a large amount of shear [46, 51], forces typically placed on the polymer during the pressurized loading process. The result is a polymer solution that provides very high-resolution separations over a short distance, which can also be loaded into a microfluidic device in a reasonable amount of time.

12.3

Conclusions

With the advances in both integrated microfluidic devices and high-performance polymeric materials discussed in this chapter, we hope that the reader has a clearer picture of where Sanger-based sequencing technologies are heading in the future. The immense potential of integration of these devices and the importance of long-read lengths for some projects in genomic sequencing signify that the Sanger approach and DNA electrophoresis will continue to be a powerful tool for future genomic technology, especially for DNA sequencing projects aimed at decoding and correctly assembling complex, repeat-rich genomes.

References

- 1 Mathies, R.A. and Huang, X.C. (1992) Capillary array electrophoresis – an approach to high-speed, high-throughput DNA sequencing. *Nature*, **359** (6391), 167–169.
- 2 Kheterpal, I. *et al.* (1996) DNA sequencing using a four-color confocal fluorescence capillary array scanner. *Electrophoresis*, **17** (12), 1852–1859.
- 3 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74** (12), 5463–5467.
- 4 Woolley, A.T. and Mathies, R.A. (1995) Ultra-high-speed DNA sequencing using capillary electrophoresis chips. *Analytical Chemistry*, **67** (20), 3676–3680.
- 5 Waters, L.C. *et al.* (1998) Microchip device for cell lysis, multiplex PCR amplification and electrophoretic sizing. *Analytical Chemistry*, **70** (1), 158–162.
- 6 Liu, S.R. *et al.* (2000) Automated parallel DNA sequencing on multiple channel microchips. *Proceedings of the National Academy of Sciences of the United States of America*, **97** (10), 5369–5374.
- 7 Salas-Solano, O. *et al.* (2000) Optimization of high-performance DNA sequencing on short microfabricated electrophoretic devices. *Analytical Chemistry*, **72** (14), 3129–3137.
- 8 Shi, Y.N. (2006) DNA sequencing and multiplex STR analysis on plastic microfluidic devices. *Electrophoresis*, **27** (19), 3703–3711.
- 9 Jacobson, S.C. *et al.* (1994) High-speed separations on a microchip. *Analytical Chemistry*, **66** (7), 1114–1118.
- 10 Blazej, R.G., Kumaresan, P. and Mathies, R.A. (2006) Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **103** (19), 7240–7245.
- 11 Paegel, B.M., Yeung, S.H.I. and Mathies, R.A. (2002) Microchip bioprocessor for integrated nanovolume sample

- purification and DNA sequencing. *Analytical Chemistry*, **74** (19), 5092–5098.
- 12 Schilling, E.A., Kamholz, A.E. and Yager, P. (2002) Cell lysis and protein extraction in a microfluidic device with detection by a fluorogenic enzyme assay. *Analytical Chemistry*, **74** (8), 1798–1804.
- 13 Andersson, H., van der Wijngaart, W. and Stemme, G. (2001) Micromachined filter-chamber array with passive valves for biochemical assays on beads. *Electrophoresis*, **22** (2), 249–257.
- 14 Broyles, B.S., Jacobson, S.C. and Ramsey, J.M. (2003) Sample filtration, concentration, and separation integrated on microfluidic devices. *Analytical Chemistry*, **75** (11), 2761–2767.
- 15 Oleschuk, R.D. *et al.* (2000) Trapping of bead-based reagents within microfluidic systems: on-chip solid-phase extraction and electrochromatography. *Analytical Chemistry*, **72** (3), 585–590.
- 16 Brody, J.P. and Yager, P. (1997) Diffusion-based extraction in a microfabricated device. *Sensors and Actuators A: Physical*, **58** (1), 13–18.
- 17 Woolley, A.T. *et al.* (1996) Functional integration of PCR amplification and capillary electrophoresis in a microfabricated DNA analysis device. *Analytical Chemistry*, **68** (23), 4081–4086.
- 18 Liu, C.N., Toriello, N.M. and Mathies, R.A. (2006) Multichannel PCR-CE microdevice for genetic analysis. *Analytical Chemistry*, **78** (15), 5474–5479.
- 19 Blazej, R.G. *et al.* (2007) Inline injection microdevice for attomole-scale Sanger DNA sequencing. *Analytical Chemistry*, **79** (12), 4499–4506.
- 20 Paegel, B.M. *et al.* (2002) High throughput DNA sequencing with a microfabricated 96-lane capillary array electrophoresis bioprocessor. *Proceedings of the National Academy of Sciences of the United States of America*, **99** (2), 574–579.
- 21 Easley, C.J. *et al.* (2006) A fully integrated microfluidic genetic analysis system with sample-in-answer-out capability. *Proceedings of the National Academy of Sciences of the United States of America*, **103** (51), 19272–19277.
- 22 Phinney, J.R. *et al.* (2004) The design and testing of a silica sol–gel-based hybridization array. *Journal of Non-Crystalline Solids*, **350**, 39–45.
- 23 Wen, J. *et al.* (2006) DNA extraction using a tetramethyl orthosilicate-grafted photopolymerized monolithic solid phase. *Analytical Chemistry*, **78** (5), 1673–1681.
- 24 Wolfe, K.A. *et al.* (2002) Toward a microchip-based solid-phase extraction method for isolation of nucleic acids. *Electrophoresis*, **23** (5), 727–733.
- 25 Legendre, L.A. *et al.* (2006) A simple, valveless microfluidic sample preparation device for extraction and amplification of DNA from nanoliter-volume samples. *Analytical Chemistry*, **78** (5), 1444–1451.
- 26 Easley, C.J., Humphrey, J.A.C. and Landers, J.P. (2007) Thermal isolation of microchip reaction chambers for rapid non-contact DNA amplification. *Journal of Micromechanics and Microengineering*, **17** (9), 1758–1766.
- 27 Roper, M.G. *et al.* (2007) Infrared temperature control system for a completely noncontact polymerase chain reaction in microfluidic chips. *Analytical Chemistry*, **79** (4), 1294–1300.
- 28 Easley, C.J., Karlinsey, J.M. and Landers, J.P. (2006) On-chip pressure injection for integration of infrared-mediated DNA amplification with electrophoretic separation. *Lab on a Chip*, **6** (5), 601–610.
- 29 Chaisson, M., Pevzner, P. and Tang, H.X. (2004) Fragment assembly with short reads. *Bioinformatics*, **20** (13), 2067–2074.
- 30 Warren, R.L. *et al.* (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **23** (4), 500–501.
- 31 Salas-Solano, O. *et al.* (1998) Routine DNA sequencing of 1000 bases in less than one hour by capillary electrophoresis with replaceable linear polyacrylamide

- solutions. *Analytical Chemistry*, **70** (19), 3996–4003.
- 32 Buchholz, B.A. *et al.* (2001) Microchannel DNA sequencing matrices with a thermally controlled “viscosity switch”. *Analytical Chemistry*, **73** (2), 157–164.
 - 33 Doherty, E.A.S. *et al.* (2002) Critical factors for high-performance physically adsorbed (dynamic) polymeric wall coatings for capillary electrophoresis of DNA. *Electrophoresis*, **23** (16), 2766–2776.
 - 34 Fredlake, C.P.H., Kan, D.G., Chiesl, C.W., Root, T.N., Forster, B.E., Barron, R.E. and Ultrafast, A.E. (2008) DNA sequencing on a microchip by a hybrid separation mechanism that gives 600 bases in 6.5 minutes. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 476–481.
 - 35 Shi, Y.N. and Anderson, R.C. (2003) High-resolution single-stranded DNA analysis on 4.5 cm plastic electrophoretic microchannels. *Electrophoresis*, **24** (19–20), 3371–3377.
 - 36 Zhou, H.H. *et al.* (2000) DNA sequencing up to 1300 bases in two hours by capillary electrophoresis with mixed replaceable linear polyacrylamide solutions. *Analytical Chemistry*, **72** (5), 1045–1052.
 - 37 de Carmejane, O. *et al.* (2001) Three-dimensional observation of electrophoretic migration of dsDNA in semidilute hydroxy-ethylcellulose solution. *Electrophoresis*, **22** (12), 2433–2441.
 - 38 Chiesl, T.N., Forster, R.E., Root, B.E., Larkin, M. and Barron, A.E. (2007) Stochastic single-molecule videomicroscopy methods to measure electrophoretic DNA migration modalities in polymer solutions above and below entanglement. *Analytical Chemistry*, **79**, 7740–7747.
 - 39 Ogston, A.G. (1958) The spaces in a uniform random suspension of fibres. *Transactions of the Faraday Society*, **54** (11), 1754–1757.
 - 40 Lunney, J., Chrambach, A. and Rodbard, D. (1971) Factors affecting resolution, band width, number of theoretical plates, and apparent diffusion coefficients in polyacrylamide gel electrophoresis. *Analytical Biochemistry*, **40** (1), 158–173.
 - 41 Kantor, R.M. *et al.* (1999) Dynamics of DNA molecules in gel studied by fluorescence microscopy. *Biochemical and Biophysical Research Communications*, **258** (1), 102–108.
 - 42 Sartori, A., Barbier, V. and Viovy, J.L. (2003) Sieving mechanisms in polymeric matrices. *Electrophoresis*, **24** (3), 421–440.
 - 43 Barron, A.E., Sunada, W.M. and Blanch, H.W. (1996) Capillary electrophoresis of DNA in uncrosslinked polymer solutions: evidence for a new mechanism of DNA separation. *Biotechnology and Bioengineering*, **52** (2), 259–270.
 - 44 Barron, A.E., Blanch, H.W. and Soane, D.S. (1994) A transient entanglement coupling mechanism for DNA separation by capillary electrophoresis in ultradilute polymer solutions. *Electrophoresis*, **15** (5), 597–615.
 - 45 Barron, A.E., Soane, D.S. and Blanch, H.W. (1993) Capillary electrophoresis of DNA in uncross-linked polymer solutions. *Journal of Chromatography A*, **652** (1), 3–16.
 - 46 Chiesl, T.N. *et al.* (2006) Self-associating block copolymer networks for microchip electrophoresis provide enhanced DNA separation via “inchworm” chain dynamics. *Analytical Chemistry*, **78** (13), 4409–4415.
 - 47 McCormick, C.L., Nonaka, T. and Johnson, C.B. (1988) Water-soluble copolymers. 27. Synthesis and aqueous solution behavior of associative acrylamide *N*-alkylacrylamide copolymers. *Polymer*, **29** (4), 731–739.
 - 48 Hill, A., Candau, F. and Selb, J. (1993) Properties of hydrophobically associating polyacrylamides – influence of the method of synthesis. *Macromolecules*, **26** (17), 4521–4532.
 - 49 Biggs, S., Selb, J. and Candau, F. (1992) Effect of surfactant on the solution properties of hydrophobically modified

- polyacrylamide. *Langmuir*, **8** (3), 838–847.
- 50 Biggs, S., Selb, J. and Candau, F. (1993) Copolymers of acrylamide/*N*-alkylacrylamide in aqueous solution – the effects of hydrolysis on hydrophobic interactions. *Polymer*, **34** (3), 580–591.
- 51 Volpert, E., Selb, J. and Candau, F. (1998) Associating behaviour of polyacrylamides hydrophobically modified with dihexylacrylamide. *Polymer*, **39** (5), 1025–1033.
- 52 Kujawa, P. *et al.* (2003) Compositional heterogeneity effects in multisticker associative polyelectrolytes prepared by micellar polymerization. *Journal of Polymer Science, Part A: Polymer Chemistry*, **41** (21), 3261–3274.
- 53 Volpert, E., Selb, J. and Candau, F. (1996) Influence of the hydrophobe structure on composition, microstructure, and rheology in associating polyacrylamides prepared by micellar copolymerization. *Macromolecules*, **29** (5), 1452–1463.

Part Five

Next-Generation Sequencing: Truly Integrated Genome Analysis

13

Multiplex Sequencing of Paired End Ditags for Transcriptome and Genome Analysis

Chia-Lin Wei and Yijun Ruan

13.1

Introduction

With the complete human genome sequence in hand [1, 2], we are now on the verge of identifying all functional genetic elements encoded in the human genome [3], and begin to elucidate the complex regulatory networks that coordinate functions of all genetic elements [4]. Furthermore, with the availability of the reference and a number of individual human genome sequences, we are able to analyze the differences of genetic variations in our entire genetic makeup to understand what genetic variations determine our susceptibility to diseases and response to drug treatments. All these promises, in fact, require much more advanced sequencing capability to rapidly decode millions or billions of DNA base pair with high efficiency.

The recently developed next-generation DNA sequencing technologies, represented by 454 and Solexa platforms [5, 6], have triggered a new wave of revolution in many aspects of genomics and biological research. A number of other DNA sequencing platforms are also on their way to the marketplace, including ABI's SOLiD system using ligation-based sequencing method [7] and Helicos' single DNA molecule sequencing platform [8]. Although these new sequencing systems are, in fact, quite different from one another and at various stages of maturity, the major advantage of these new sequencing technologies is their ability to generate large volumes of DNA sequence data, from hundreds of megabases to several billion bases per machine run without requiring tedious DNA cloning and labor-intensive preparation of sequencing templates. However, the obvious weak point shared by all of the current sequencing technologies is the short read length (100–200 bp by 454s GSFLX; 25–35 bp by Solexa, SOLiD, or Heliscope). Despite this limitation, the highly multiplex nature of these new sequencing methods has already made a tremendous impact.

An immediately and widely recognized solution to overcome the problems associated with short tag-based sequencing platforms is to adapt the paired end ditag (PET) sequencing strategy that we had originally developed for cloning-based transcriptome analysis [9] and whole genome mapping of transcription factor

binding sites (TFBS) [10]. We have demonstrated that the PET scheme can be easily adapted to the tag-based multiplex sequencing platform [11], in which PET can overcome for the inherent limitations of short reads by providing paired end information from long contiguous DNA fragments. The multiplex sequencing of paired end ditags (MS-PET) not only extends the linear DNA sequence coverage but, more importantly, also enables one to infer the relationship between the two ends of DNA fragments in defined distance and content. This unique feature has enabled us to identify unconventional fusion transcripts [12] and genome structural variations (SVs) [13]. Therefore, the MS-PET sequencing strategy can not only offer a unique advantage to improve the efficiency and accuracy of short tag-based sequencing methods, but can also expand their applications and information outputs.

Here, we describe the concept of the MS-PET approach and its applications in transcriptome analysis, mapping of transcription factor binding sites, characterization of long range chromatin interactions, and identification of genome structural variations. The potential application of PET analysis in resequencing the human genome and its implication in personalized genome medicine are also discussed. Collectively, the PET technology discussed in this chapter has a tremendous and immediate value in providing unique and comprehensive solutions in this fast-growing research area for whole genome characterization of transcription and epigenetic regulatory networks and whole genome scan for chromosomal structural variations.

13.2

The Development of Paired End Ditag Analysis

Our interest started with the use of full-length cDNA cloning and sequencing [14] and SAGE/MPSS [15–17] approaches to characterize mammalian transcriptomes. However, we realized that the full-length cDNA approach was too expensive and labor intensive, while the short-tag approaches, although efficient in counting cDNA tags, could not specify transcripts in terms of where they start (5' end) and where they terminate (3' end). To improve our capability for comprehensive transcriptome analysis, we first developed 5'LongSAGE and 3'LongSAGE protocols to map transcription start sites (TSS) and polyadenylation sites (PAS) of gene [18]. Expanding from such capability, we then devised the paired end ditagging strategy, in which the 5' and 3' tags derived from a DNA fragment are covalently linked as a single molecule for sequencing analysis.

The principal concept of the PET methodology is to extract the paired end signatures from each of the target DNA fragments and map the paired tag sequences to the reference genome for accurate demarcation of the boundaries of the tested DNA fragments in the genome landscape. Starting from any kind of DNA (cDNA or genomic DNA), in this process, specific adapter sequences are ligated to the DNA fragments, and the ligated DNAs are then digested by type IIs restriction enzymes to release the paired end ditags, with tag from each end of 20 bp in length. The connectivity between the two paired tags is achieved through the use of cloning

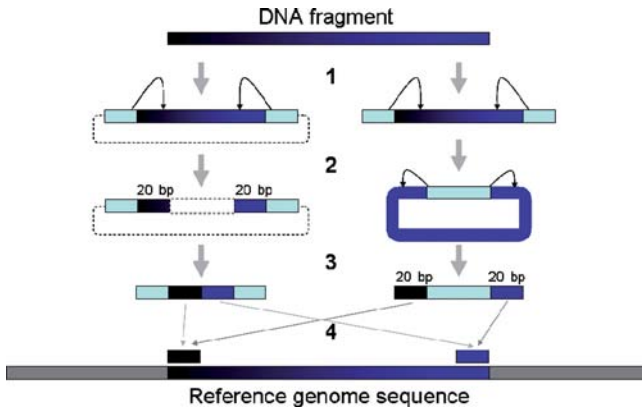


Figure 13.1 The schematic view of paired end ditag analysis. *Left*: cloning-based PET method. Adapter-ligated DNA fragments are cloned into the vector before they are subjected to type II restriction enzyme digestion. *Right*: the cloning-free PET approach. Adapter-ligated DNA fragments are self-circularized followed by type II restriction enzyme digestions. The resulted PETs are sequenced and map to genome to determine the identity of DNA fragments of interest.

vector that embraces the insert DNA fragment [9]. Later, we simplified this step by circularization of the adapted DNA fragments (the *in vitro* cloning-free method) (Figure 13.1). The PET structure containing two paired short tags can then be either concatenated into longer stretch of DNA fragments for efficient high-throughput sequencing by traditional capillary method, in which each sequencing read can reveal 10–20 PET sequence units, or directly sequenced by the multiplex sequencing method. The PET sequences are mapped to the reference genome sequences. The paired 5' and 3' signatures were considered mapped if they were located on the same chromosome, same strand (+ or –), and within the expected genomic distance of each other. As a result of such mapping criteria, the vast majority of these PETs can be uniquely located to the reference genome and can accurately define the identity of the DNA fragments analyzed.

Based on the PET concept, we first developed the Gene Identification Signature (GIS) analysis using paired end ditags (GIS-PET) to analyze full-length cDNA fragments and transcriptome. The full-length cDNA fragments are first cloned into vector as full-length cDNA library, then the cDNA inserts are digested by MmeI restriction enzyme, and the two tags remained on the cloning vector are jointed through circularization ligation reaction. The resulting single PET constructs are concatenated into longer DNA fragments for high-throughput sequencing by traditional capillary method. The PET sequences can precisely demarcate the boundaries of full-length transcripts on genome landscape [9]. Immediately afterward, we invented the ChIP-PET (chromatin immunoprecipitation coupled with paired end ditagging) analysis for highly accurate, robust, and unbiased genome-wide identification of transcription factor binding sites [10]. When the 454 sequencing technology became available, we immediately adapted the new sequencing

method and developed the multiplex sequencing for paired end ditag analysis [11], which achieved an additional 100-fold efficiency compared to that of conventional sequencing method used for PET experiments.

13.3

GIS-PET for Transcriptome Analysis

Traditionally, transcripts and transcriptomes are studied by DNA microarrays and cDNA sequencing. The advancement of DNA microarray fabrication to cover the entire genome has provided an attractive approach to study transcriptome. The genome-wide tiling arrays [19] provided a massive parallel approach for characterization of all expressed exons and a promise for highly comprehensive transcriptome analysis. However, the tiling array data have no inherent structural information for each transcript to be characterized; that is, they are not straightforward to define the start and termination positions of individual transcript units and the connectivity of each exons. Furthermore, the tiling array approach suffers from cross-hybridization noise when it is used to detect transcripts expressed in highly homologous genomic regions. In contrast, the sequencing-based strategies, such as full-length cDNA sequencing [20] and short tag sequencing of SAGE [15, 16] and MPSS [17] had contributed immensely to transcriptome data, but were limited by huge operational cost, inefficiency (full-length cDNA sequencing), or insufficient information (SAGE and MPSS tags) (Figure 13.2).

This PET-based sequencing approach can accurately demarcate the individual transcripts of different alternative forms expressed. Furthermore, this property enables the identification of unconventional transcripts such as those formed by intergenic, bicistronic linkage, or transsplicing events, which can be uncovered by using the array-based approach. In GIS-PET analysis, the transcriptome of a biological sample is first constructed as a full-length cDNA library that captures all intact transcripts. This library is then subjected for PET analysis. PETs from the two ends of each expressed full-length transcript (18 bp from 5' end and 18 bp from 3' end) are extracted and subjected to multiplex sequencing (Figure 13.2). Millions of transcripts represented by PETs are analyzed and the PET sequences are precisely mapped to the genome for the identification of expressed genes and quantification of expression levels. We have demonstrated that over 98% of PET sequences are precise in demarcating the 5' end and the 3' end of full-length transcripts, and the copy numbers of PETs mapping to specific loci provide digital counts of gene expression level. In addition to analyzing expressed transcriptomes and accurately demarcating gene transcription boundaries, GIS-PET can also infer proximal promoter sites and enable the discovery of novel genes and alternative transcript variants with unprecedented efficiency. These efficient features make GIS-PET as the ideal approach for gene identification and differential quantification. We have successfully applied GIS-PET in the FANTOM project to conduct a comprehensive mouse transcriptome analysis [21] and in the ENCODE project to characterize the 1% human genome for functional DNA elements [4].

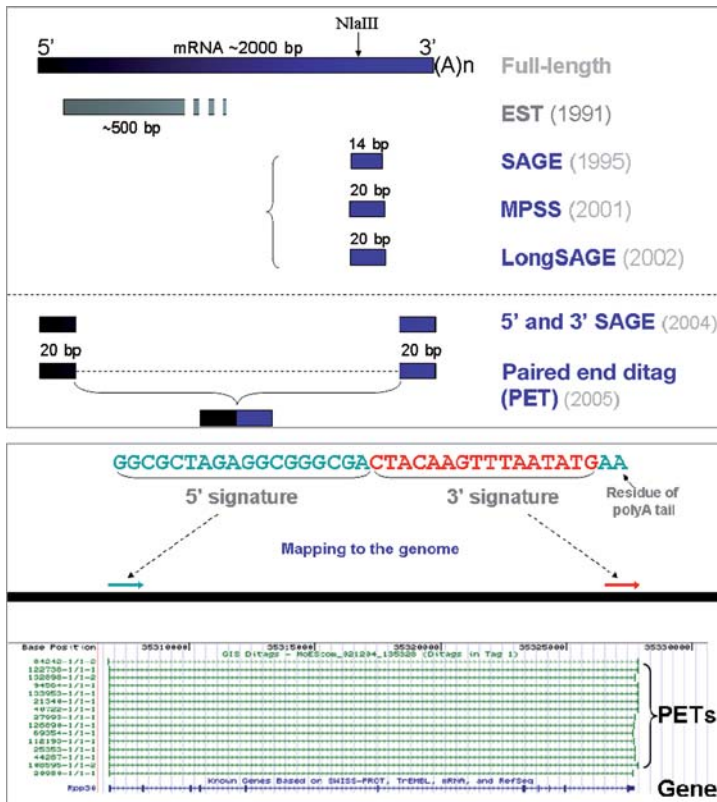
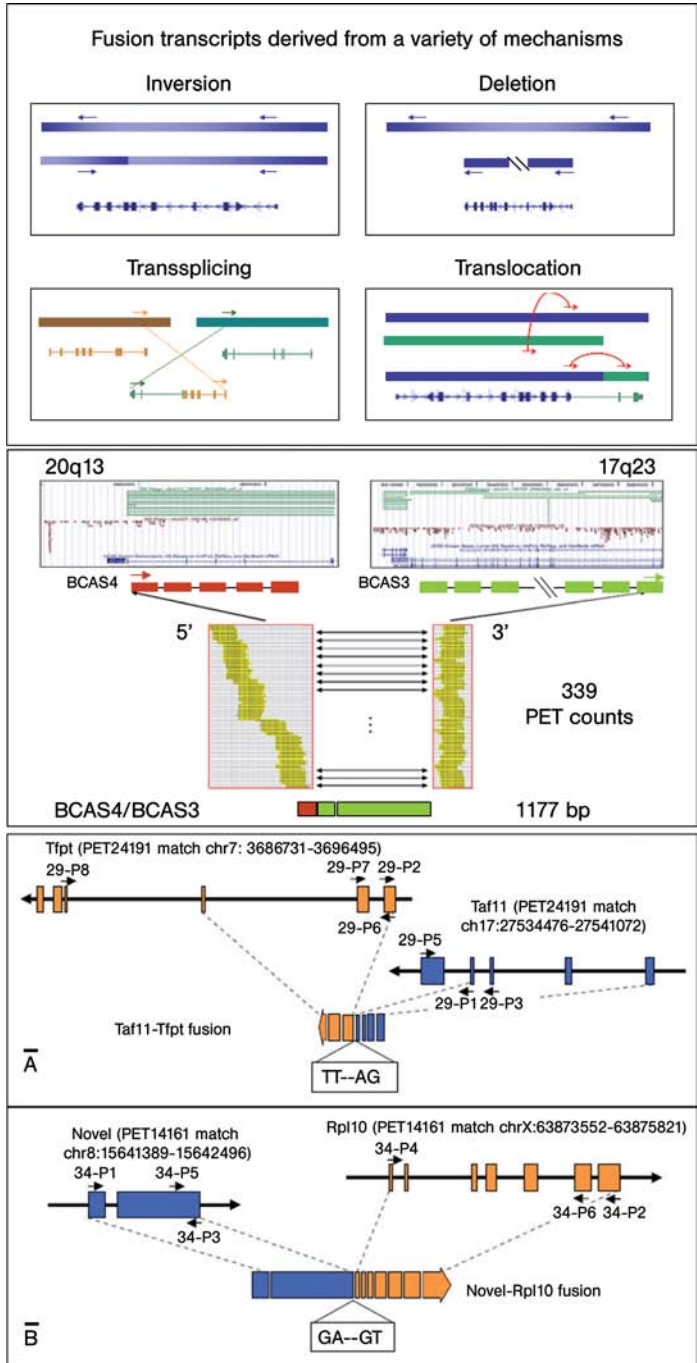


Figure 13.2 GIS-PET for transcriptome analysis. *Top:* Diagram describes various sequencing strategies used for transcript identification and profiling, when they became available, and the information content these methods provided. Paired end ditags were developed by connecting the 5' and 3' tags from each end of full-length cDNA fragments. *Bottom:* Example of a PET structure, 5' and 3' tags mapped to reference genome to define the TSS and PAS of expressed transcripts and the number of PETs representing transcript abundance.

The most unique feature of GIS-PET is its ability to delineate the relationship between the two ends of individual cDNA molecule. GIS-PET can efficiently discover fusion genes resulting from genome rearrangements or transsplicing (Figure 13.3). Specifically, fusion genes resulted from PETs located on different orientations or different chromosomes can be identified through PET mapping. It is known that there are a variety of mechanisms that generate fusion transcripts with novel functions, and fusion genes have been shown as a valuable tool for tumor diagnosis and therapeutic stratifications [22]. *BCR-ABL* translocation, the well-known example, was used successfully in the discovery of new diagnostic marker and the development of Gleevec in CML [23]. We have applied the GIS-PET to two of the well-studied cancer cell lines, breast cancer MCF7 and colon cancer HCT116, and successfully identified more than 70 potential fusion genes [12]. One of them, *BCAS4/BCAS3*, has been verified through independent cytogenetic and



genomic DNA sequencing approaches (Figure 13.3). We have also identified a number of fusion genes derived from transsplicing mechanism in mouse embryonic stem cells [9]. Our data suggest that there are significant numbers of fusion genes that may have important biological functions and yet to be identified in many different cell types, including stem and cancer cells. Indeed, GIS-PET is the only efficient system for large-scale discovery of this uncharted territory. By coupling the GIS-PET method with the next-generation DNA sequencing platforms, we can set up a large-scale program to specifically screen unconventional fusion transcripts derived from various mechanisms.

13.4

ChIP-PET for Whole Genome Mapping of Transcription Factor Binding Sites and Epigenetic Modifications

It is known that gene transcription in eukaryotic cells is regulated by specific transcription factors with specific DNA recognition properties through direct or indirect binding to the regulatory DNA elements. Thus, the identification of functional elements such as transcription factor binding sites on a whole genome level is the next challenge for genome sciences and gene regulation studies. Increasing evidence also suggests that many TFs function in a cooperative manner to form complex transcription regulatory circuits. However, studying such complex regulatory networks has been a big challenge of biology. These fundamental questions can be addressed through whole genome comprehensive profiling of the transcription factor DNA interactions and characterization of chromatin structures mediated by specific transcription factor interactions. The most widely used method for transcription factor binding sites had been ChIP-chip [24], in which living cells were fixed with formaldehyde that cross-link the DNA/protein interactions *in vivo*. After fragmentation, the chromatin complexes were bound by specific antibody against given protein factor and therefore enriched the target DNA fragments associated with TF. The enriched DNA fragments were then detected by DNA microarray [25]. Owing to the large size and complexity of mammalian genomes, the DNA microarrays constructed often contained partial genome information or only promoter regions of well-characterized genes. Therefore, many of the ChIP-chip analyses were in fact incomplete.

To profile transcription factor binding in an unbiased fashion and genome-wide, we devised the ChIP-PET method using paired end ditag sequencing to characterize

Figure 13.3 Unconventional fusion transcripts identified by GIS-PET analysis. *Top*: Types of fusion transcripts can be identified by GIS-PET analysis and the resulted PETs mapping patterns on the reference genome. *Middle*: BCAS4/BCAS3 fusion genes uncovered by GIS-PET analysis from MCF7 cells. Multiple PET clusters (339 PETs) with their 5' tags mapped to chr20q13, the starting region of BCAS4 gene, whereas the 3' tags mapped to chr17q23, the 3' terminal region of BCAS3 gene. The resulted fusion transcript is 1177 bp. *Bottom*: Two different fusion genes derived from transsplicing events in mouse embryonic stem cells.

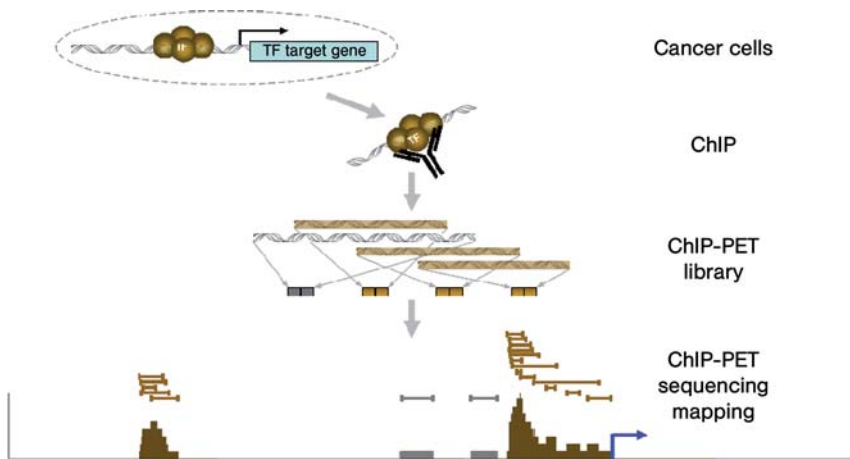


Figure 13.4 ChIP-PET to map transcription factor binding sites. DNA fragments bound by specific transcription factor are enriched by chromatin immunoprecipitation. PETs from these DNA can be extracted, sequenced, and mapped to the genome. The genomic loci shared by multiple PET mapping regions can be inferred as TFBS.

the ChIP-enriched DNA fragments [10]. In the ChIP-PET method, ChIP-enriched DNA fragments are extracted for the paired 5' and 3' tags. These paired end ditags are subjected for ultrahigh-throughput sequencing [11], and the PET sequences are accurately mapped to reference genome for demarcating the locations of inferred ChIP DNA fragments. The genuine transcription factor binding sites can be identified through overlapping of PET-inferred ChIP DNA fragments (Figure 13.4).

The ChIP-PET method has proved that high-throughput sequencing is a superior readout to identify TFBS compared to the ChIP-chip approach. We have demonstrated that more than 99% of the binding loci determined by PET clustering can be verified by ChIP-qPCR validation experiments, and the PET-defined binding regions can be narrowed down to less than 10 bp [10]. Using ChIP-PET, we have successfully mapped the whole genome binding profiles for a number of important transcription factors, including p53 [10], Oct4 and Nanog [26], cMyc [27], ER α [28], and NF- κ B [29]. We have also mapped the epigenomic profiles of histone modifications in human embryonic stem cells [30].

Encouraged by the ChIP-PET results, the sequencing-based approach for TFBS analysis was elevated to a new level in 2007 when the Solexa sequencing platform became available. The advantages of Solexa sequencing in ChIP analysis are its massive output of short tag sequencing reads (>40 million tag reads of 25 bp per machine run) and its requirement for small amount of input DNA (ng level). These advantages have been promptly appreciated and recognized as a new method ChIP-seq, in which ChIP DNA fragments are directly sequenced to generate millions of short tags to identify TFBS. The application of ChIP-seq has generated exciting

results in mapping histone modifications [6, 31] and TFBS [32]. These reports have demonstrated that ChIP-seq is a simple and robust approach for global profiling of the transcription factor binding sites with high specificity and accuracy. In summary, sequencing-based approach is a superior approach for whole genome protein/DNA interaction analysis, and the accumulation of such results will be useful to construct the systematic transcriptional networks and regulatory circuitries. Such global approach should provide invaluable information to decipher the gene regulation programs.

13.5

ChIA-PET for Whole Genome Identification of Long-Range Interactions

As discussed above, significant progress has been made in identifying genes and regulatory elements that modulate transcription and replication of the genome. A growing body of data generated by us [10, 26, 27] and others [33, 34] has started to show that a large portion of the putative regulatory elements are localized far away from genes coding regions; this phenomenon is difficult to explain by the simple linear relationship of locations of such elements along the genome.

It has been hypothesized for years that through tertiary conformation of chromatin, DNA elements can function at considerable genomic distance from the genes they regulate [35]. Studies of *in vivo* chromatin interactions in specific cases have demonstrated that long-range (up to megabase scale) enhancers and locus control regions (LCRs) can be found in close spatial proximity to their target genes [36, 37]. Emerging data from recent studies also raise the possibility that chromosomes can interact with each other to regulate transcription in *trans* [38], and suggest that higher level interactions of DNA elements in nuclear three-dimensional space are important and abundant mechanisms for the regulation of genome functions.

Most of the recent advances in the understanding of chromatin interactions have relied on the chromosomal conformation capture (3C) method [39], which has proven very useful in determining the identity of DNA sequences from remote genomic locations that lie in close physical proximity in formaldehyde cross-linked chromatin. However, this method is limited to the detection of only specific interactions for which we have prior knowledge or perception of their existence. To overcome this limitation, a number of groups have developed chromosome conformation capture on chip or by cloning, 4C [40], and 5C [41] methods to expand the scope for the detection of such chromatin interactions (Figure 13.5). Based on 3C-circularized ligation products, the 4C approach uses PCR to prime on known targets and subsequently extends into DNA fragments of unknown regions. The 4C-amplified products can then be characterized either by microarrays or by cloning and sequencing analysis. Hence, the 4C method has the potential to detect many chromatin interactions *de novo* from a known site (Figure 13.5). Also starting from the 3C ligation products, the 5C method uses multiplex oligonucleotide primers predesigned around many locations of the restriction sites used for 3C analysis over a large genomic region, and uses ligation-mediated amplification (LMA) to amplify the

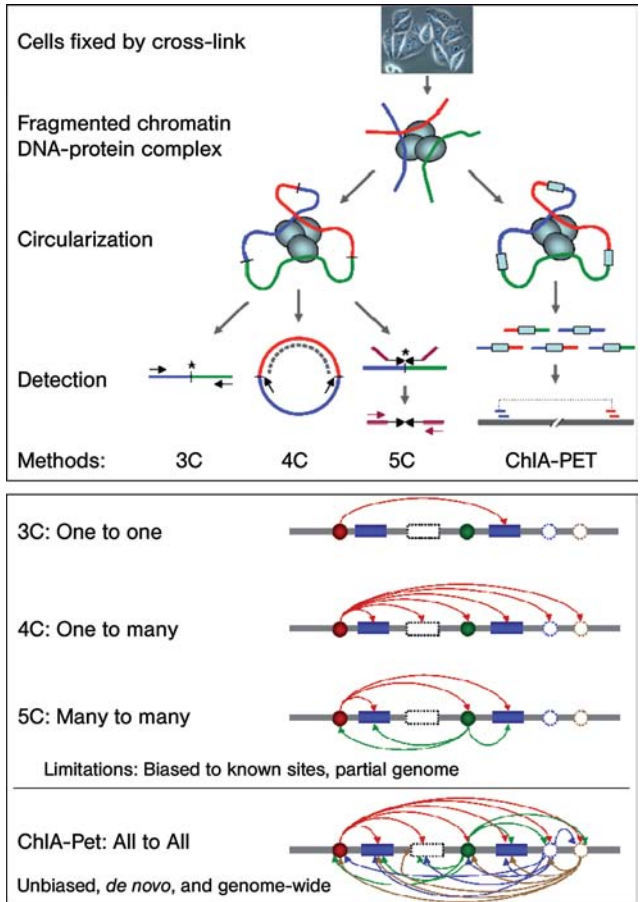


Figure 13.5 Illustration of different approaches used to characterize long-range chromatin interactions. *Left:* Tethered DNA bound by specific protein factors are ligated. Chromatin interactions can only be detected by PCR using specific PCR primers against the known regions (3C), sequencing analysis of DNA regions from primers extended from specific regions (4C), or ligation-mediated PCR from specific designed PCR primers (5C). *Right:* In ChIA-PET analysis, adapters containing type IIa RE are ligated at each end of tethered DNA. PETs from interacting chromatin DNAs are generated by RE digestion (MmeI or EcoPI5I) and sequenced to determine the interactions at genome-wide scale. *Below:* the scale, resolution, and efficiency of ChIA-PET analysis compared with other similar approaches.

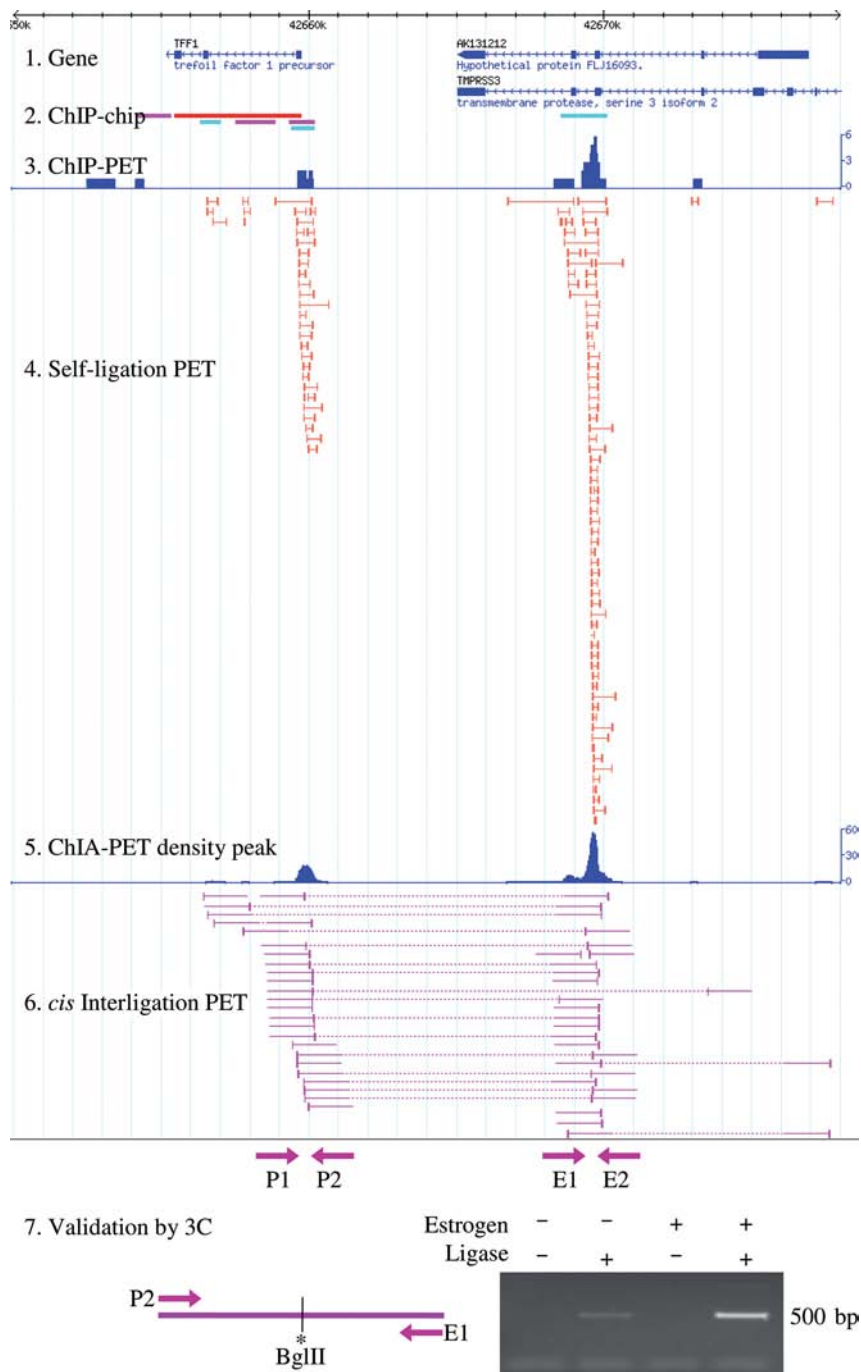
3C products. The 5C amplicons can then be analyzed by specifically designed microarrays or DNA sequencing. The 5C primers are based on restriction sites instead of known targets; therefore, it has the potential to detect chromatin interactions *de novo* (Figure 13.5). However, the current version of 5C is constrained by the limits of reliable LMA, and thus can be used only for partial genome. There is no doubt that the 4C and 5C methods will trigger a new wave of interest into the

long-range control of transcription regulation. However, these methods essentially rely on the same concept of PCR selection for targeted interactions and are therefore unable to achieve an unbiased whole genome analysis for the *de novo* discovery of chromatin interactions. To move this field further forward into three-dimensional space, it is desirable to develop an unbiased, whole genome approach that is independent of any prior knowledge or information about any sequence feature for the *de novo* discovery of chromatin interactions, especially those involved in transcriptional regulation.

We therefore devised the chromatin interaction analysis by using paired end ditag (ChIA-PET) (Figure 13.5), a genome-wide, high-throughput, unbiased, and *de novo* approach for detecting long-range chromatin interactions in 3D nuclear space. This method combines the “proximity ligation” principle used in the 3C approach [35] with the power of PETs [5] and next-generation sequencing technologies [3, 37]. Briefly, a specially designed DNA oligonucleotide sequence is introduced to link different DNA fragments that are nonlinearly related in the genome but brought together in close spatial proximity by protein factors *in vivo*. The proximity ligation products are then digested by designated type IIs restriction enzyme (MmeI or EcoP15I) to release paired end ditag structures from each of the ligated DNA fragments. The ditags are subjected for high-throughput sequencing, and the PET sequences are mapped to the reference genome to reveal the relationship between the paired DNA fragments and to identify long-range interactions among DNA elements. Therefore, by design, the ChIA-PET approach is for detection of all interactions in a system, while the 3C method is for detection of only one-to-one interactions and 4C is for one-to-many (Figure 13.5).

We have begun to demonstrate this approach in the model yeast and in the human genome. Through our ChIA-PET experiments that characterized the estrogen receptor (ER α)-mediated interactions in human breast adenocarcinoma cells (Figure 13.6), we have shown that ChIA-PET can provide two types of information: the self-ligation PETs that define the transcription factor binding sites and the interligation PETs that reveal the interactions between the binding sites. In this work, we provided the first global view of long-range chromatin interactions mediated by transcription factors in the human genome. Our data have also provided evidence that multiple looping structure of long-range interactions is a primary mechanism for transcription regulation by ER α , which may facilitate the recruitment of collaborative cofactors and basal transcription components to the desired target sites.

These results convincingly demonstrated that ChIA-PET is an unbiased, *de novo*, and high-throughput approach for the global analysis of long-range chromatin interactions mediated by protein factors. With the availability of the tag-based ultrahigh-throughput sequencing technologies, ChIA-PET has the potential to become the most versatile and informative technology to map transcription interactomes mediated by all transcription factors and chromatin-remodeling proteins. Such whole genome binding and interaction information will certainly open a new field for understanding transcription regulation mechanisms at three-dimensional levels.



13.6 Perspective

As exemplified in the above sections, PET has been well established as a primary technology for the characterization of transcriptome and elucidation of transcription regulatory networks. In addition, because of its unique ability to derive the relationship between the two ends of test DNA fragments, it has been perceived to have great application potential in genomic analysis, such as the analysis of genome structural variations, as well as *de novo* individual and personal genome shotgun sequencing and assembling.

Growing evidence has shown that human genomes undergo substantial structural variations including large and small pieces of insertion, inversion, translocation, and copy number variations (amplification or deletion). Traditional methods of detecting structural variations by array CGH and DNA FISH have low sensitivity and specificity. The combination of PET analysis and the next-generation sequencing offers us new promises for a comprehensive genome structural variation analysis. Recent study has demonstrated such potential by using the 454 sequencing platform and paired end sequencing of 3 kb genomic DNA fragments from two human cell cultural lines [13] and identified extensive structural variations in the human genome. Owing to abundant and large (>5 kb) repetitive regions scattered in the human genome, the current study with the ability of paired end sequencing covering only 2–3 kb genomic span would not be sufficient for a comprehensive survey of human genome structural variations. It will be desirable to develop the capability of paired end sequencing to cover a genomic span of more than 10 kb.

Ultimately, with further improvement of the PET technology particularly the robust preparation of PET libraries from large genomic DNA fragments (at 10, 20, or 50 kb) and maturation of the ultrahigh-throughput tag-based sequencing that could generate billions of tag reads per run, it is possible that the PET-based shotgun whole genome sequencing and assembling would become the method of choice for *de novo* personal human genome sequencing.

Figure 13.6 ER α ChIA-PET data mapped at the TFF1 locus. The ChIA-PET sequences were mapped to the reference genome for identification of ER α -binding sites and ER α -mediated chromatin long-range interactions. The self-ligation PETs would indicate ER α -binding site density, while the interligation PETs would identify long-range interactions between two DNA fragments. The tracks of information included here are (starting from the top) (1) UCSC-known genes, the TFF1 gene locus. TFF1 is an estrogen-upregulated gene in MCF7 cells; (2) ER α ChIP-chip (blue bars); (3) ER α ChIP-PET density histogram showing ER α -binding site peaks; (4) Self-ligation ChIA-PET data that show each PET with head and tail tags together with a solid horizontal line (orange) to represent the virtual ChIP DNA fragment; (5) ER α ChIA-PET density histogram showing the ER α -binding sites in this region; (6) Interligation ChIA-PET data show each tag with a vertical line and a solid horizontal line to represent the virtual DNA fragment from which the PET was derived from. A dotted line connects the paired tags of the same PET, indicating the interaction of the two DNA fragments; and (7) Interactions identified by ChIA-PET data in this dataset were validated by ChIP-3C experiments. Negative controls of [estrogen –] and [ligation –] were included. Only [estrogen +] and [ligation +] reactions provided positive ChIP-3C products.

References

- 1 International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**: 931–945.
- 2 Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biology*, **5**, e254.
- 3 The ENCODE consortium (2004) The ENCODE (ENCyclopedia Of DNA Q2 Elements) Project. *Science*, **306**, 636–640.
- 4 Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
- 5 Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- 6 Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
- 7 Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
- 8 Harris T.D., Buzby P.R., Babcock H., Beer E., Bowers J., Braslavsky I., Causey M., Colonell J., Dimeo J., Efcavitch J.W. *et al.* (2008) Single-molecule DNA sequencing of a viral genome. *Science*, **320**, 106–109.
- 9 Ng, P., Wei, C.L., Sung, W.K., Chiu, K.P., Lipovich, L., Ang, C.C., Gupta, S., Shahab, A., Ridwan, A., Wong, C.H. *et al.* (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods*, **2**, 105–111.
- 10 Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z. *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**, 207–219.
- 11 Ng, P., Tan, J.J., Ooi, H.S., Lee, Y.L., Chiu, K.P., Fullwood, M.J., Srinivasan, K.G., Perbost, C., Du, L., Sung, W.K. *et al.* (2006) Multiplex sequencing of paired-end dtags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Research*, **34**, e84.
- 12 Ruan, Y., Ooi, H.S., Choo, S.W., Chiu, K.P., Zhao, X.D., Srinivasan, K.G., Yao, F., Choo, C.Y., Liu, J., Ariyaratne, P. *et al.* (2007) Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using paired-end dtags (PETs). *Genome Research*, **17**, 828–838.
- 13 Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
- 14 Carninci, P. and Hayashizaki, Y. (1999) High-efficiency full-length cDNA cloning. *Methods in Enzymology*, **303**, 19–44.
- 15 Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
- 16 Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002) Using the transcriptome to annotate the genome. *Nature Biotechnology*, **20**, 508–512.
- 17 Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, **18**, 630–634.

- 18 Wei, C.L., Ng, P., Chiu, K.P., Wong, C.H., Ang, C.C., Lipovich, L., Liu, E.T. and Ruan, Y. (2004) 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 11701–11706.
- 19 Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engle, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G. *et al.* (2001) Experimental annotation of the human genome using microarray technology. *Nature*, **409**, 922–927.
- 20 Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F. *et al.* (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16899–16903.
- 21 Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
- 22 Mitelman, F., Johansson, B. and Mertens, F. (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nature Genetics*, **36**, 331–334.
- 23 Mauro, M.J., O'Dwyer, M., Heinrich, M.C. and Druker, B.J. (2002) STI571: a paradigm of new agents for cancer therapeutics. *Journal of Clinical Oncology*, **20**, 325–334.
- 24 Wu, J., Smith, L.T., Plass, C. and Huang, T.H. (2006) ChIP-chip comes of age for genome-wide functional analysis. *Cancer Research*, **66**, 6899–6902.
- 25 Lee, T.I., Johnstone, S.E. and Young, R.A. (2006) Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protocols*, **1**, 729–748.
- 26 Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics*, **38**, 431–440.
- 27 Zeller, K.I., Zhao, X., Lee, C.W., Chiu, K.P., Yao, F., Yustein, J.T., Ooi, H.S., Orlov, Y.L., Shahab, A., Yong, H.C. *et al.* (2006) Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 17834–17839.
- 28 Lin, C.Y., Vega, V.B., Thomsen, J.S., Zhang, T., Kong, S.L., Xie, M., Chiu, K.P., Lipovich, L., Barnett, D.H., Stossi, F. *et al.* (2007) Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genetics*, **3**, e87.
- 29 Lim, C.A., Yao, F., Wong, J.J., George, J., Xu, H., Chiu, K.P., Sung, W.K., Lipovich, L., Vega, V.B., Chen, J. *et al.* (2007) Genome-wide mapping of RELA (p65) binding identifies E2F1 as a transcriptional activator recruited by NF-kappaB upon TLR4 activation. *Molecular Cell*, **27**, 622–635.
- 30 Zhao X.D., Han X., Chew J.L., Liu J., Chiu K.P., Choo A., Orlov Y.L., Sung W.K., Shahab A., Kuznetsov V.A., *et al.* (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell*, **1**, 286–298.
- 31 Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
- 32 Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of *in vivo* protein–DNA interactions. *Science*, **316**, 1497–1502.
- 33 Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.

- 34 Carroll, J.S., Meyer, C.A., Song, J., Li, W., Geistlinger, T.R., Eeckhoute, J., Brodsky, A.S., Keeton, E.K., Fertuck, K.C., Hall, G.F. *et al.* (2006) Genome-wide analysis of References estrogen receptor binding sites. *Nature Genetics*, **38**, 1289–1297.
- 35 Fraser, P. and Bickmore, W. (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, 413–417.
- 36 Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F. and Fraser, P. (2002) Long-range chromatin regulatory interactions *in vivo*. *Nature Genetics*, **32**, 623–626.
- 37 Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W. and Fraser, P. (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics*, **36**, 1065–1071.
- 38 Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R. and Flavell, R.A. (2005) Interchromosomal associations between alternatively expressed loci. *Nature*, **435**, 637–645.
- 39 Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
- 40 Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U. *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, **38**, 1341–1347.
- 41 Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C. *et al.* (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research*, **16**, 1299–1309.

14

Paleogenomics Using the 454 Sequencing Platform

M. Thomas P. Gilbert

14.1

Introduction

In 2005, the field of ancient DNA underwent a revolution. With the publication of 26 861 bp of 40 000 ^{14}C years old cave bear (*Ursus spelaeus*) DNA, Noonan *et al.* [1] demonstrated that the field was poised to advance into the genomic era. The enormity of this advance is easily apparent in the light of the fact that prior to this, few aDNA studies had involved the amplification, sequencing, and analysis of more than 500–1000 bp fragments of DNA sequences from individual specimens. Those that had, predominantly a handful of complete mtDNA genomes from three extinct moa species (*Emeus crassus*, *Anomalopteryx didiformis*, and *Dinornis giganticus*) and two Siberian mammoths (*Mammuthus primigenius*) [2–5], had for the most part taken years of painstaking research to assemble. While the data presented by Noonan *et al.* [1] were generated through the conventional Sanger sequencing of metagenomic libraries, the year 2005 also saw the groundbreaking publication [6] that heralded the approach of the pyrosequencing-based sequencing-by-synthesis platforms (hereafter referred to as 454 approaches/platforms). Seizing the opportunity that 454 platforms offer with regard to their capability to generate greatly increased amounts of sequence information, within months of the publication of these original studies, Poinar *et al.* [7] had reported a further advance within this nascent field, publishing 13 megabases of Siberian woolly mammoth sequence extracted from an approximately 27 740 ^{14}C years old bone. Several studies followed swiftly on the heels of this publication, demonstrating that paleogenomics was not only limited to cold-preserved samples but could also be applied to temperate preserved remains such as Neandertals [8, 9]. However, not only did these publications demonstrate that paleogenomics was at last a viable field of research, but in doing so they provided the first insights into the future challenges that the field would, and will continue to, face. In particular, although the studies demonstrated that high-throughput sequencing-by-synthesis technologies help circumvent one of the classic problems associated with studying ancient DNA – the fragmentation of template molecules through

time – they also demonstrated that other forms of DNA damage, sequencing error, and sample contamination were to cause a serious hindrance.

Owing to its nascent stage, few paleogenomic studies have been published so far, and for the most part their findings detail aspects of the methods and challenges that future studies will face. Therefore, the evolutionary and biological conclusions of the studies are at this time limited. As such, the predominant focus of this chapter will be to outline how several problems characteristic to the field of ancient DNA present significant challenges to paleogenomics. Following this, I outline potential solutions that have been, or might be, applied to the studies to increase the efficiency of paleogenomic analysis. Finally, I conclude by outlining ground research that remains to be done.

14.2

The DNA Degradation Challenge

No topic related to the field of ancient DNA can be fully appreciated without at least a loose understanding of the serious challenges that are presented in the forms of DNA degradation and contamination. During and following death, DNA within any cells faces two principal problems. First, cellular repair mechanisms cease, and with them any ability to repair the subsequent multitude of DNA damage processes that affect the molecule. Second, the cell will often go through autolysis, as otherwise regulatory cellular functions cease to exercise control. The end result of both processes is similar, and catastrophic for the DNA molecule, and thus any genetic study that requires high-quality DNA templates. DNA molecules are remarkably fragile and susceptible to damage by a number of natural processes. Although the biochemical details behind the processes vary and are irrelevant in this context, most of the processes produce the same phenotypic result – the length of intact DNA molecules available for PCR amplification and sequencing analysis rapidly decreases. A number of factors affect the rate of this decay, including temperature, proximity to free water, environmental salt content, exposure to radiation, and so on [10], and as such the rate is extremely difficult to model accurately. However, a number of studies have argued that due to its key importance in most chemical reactions, temperature plays an important role, linked exponentially to the rate of degradation (cf. [11, 12]). There are two take-home messages from this relationship. First, as temperatures increase, the rate of DNA degradation increases rapidly (and conversely, as temperature falls, degradation decreases rapidly). Second, for any particular ancient or otherwise degraded sample, the absolute quantity of PCR amplifiable DNA fragments increases rapidly as template size is decreased (Figure 14.1).

These two factors play a key role with regard to the field of paleogenomics. First, DNA degradation itself plays a naturally limiting role on what can, and cannot, be analyzed using paleogenomic approaches. As with all other sequencing-based studies, regardless of the method used, the underlying requirement is enough sequenceable template molecules (in both size and quantity) from which to generate data. Given enough time, the DNA content of all dead biological tissues will decrease

to levels where the remaining fragments are too short for any meaningful information to be recovered. Therefore, ultimately a natural limit exists beyond which no DNA sequence information can be recovered. Although a series of studies from the early 1990s (and a sporadic few since) reported the recovery of DNA from samples that are millions of years old [13–16], the results of these studies are, without doubt, derived from modern sources of DNA that contaminated the samples. The second implication is simply that, due to the temperature–degradation relationship, for any given sample age, cold preservation is more likely to provide usable genetic material than warm preservation.

14.3

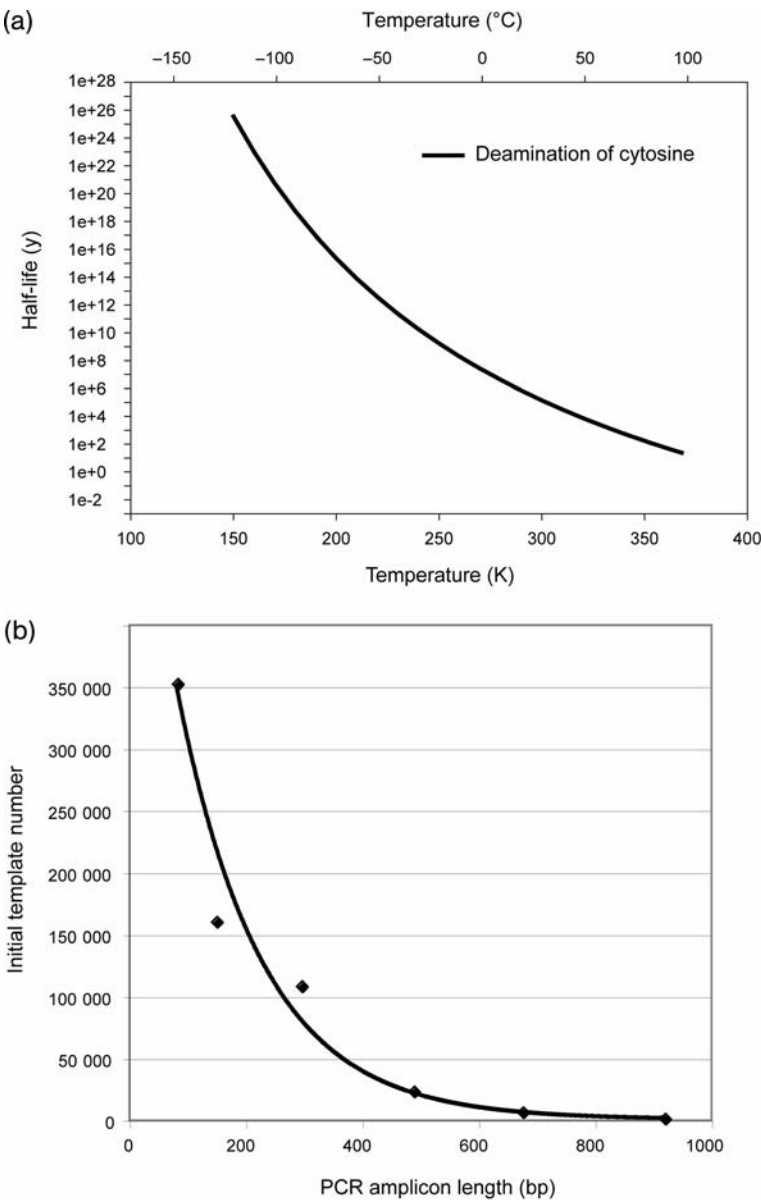
The Effects of DNA Degradation on Paleogenomics

DNA degradation has a number of direct effects on paleogenomic studies. One of the key advantages of the emulsion PCR (emPCR)/pyrosequencing-based sequencing approach of the 454 platforms over conventional methods is that they can be effectively used on much smaller template sizes. Although conventional PCR approaches require sufficient DNA template survival to allow the placement of sequence-specific primers on single undamaged DNA fragments, thus preventing amplification if the surviving fragments are below this size, the approach adopted in the 454 platforms does not require this. As such, the sole requirement in this regard is that some DNA survives that can be amplified by emPCR. It is precisely for this reason that a large amount of nuclear DNA (nuDNA) has been generated from Neandertal bone samples [8, 9] that contained only short fragments (e.g., <100 bp molecules of mtDNA) and were not regarded as having any conventional PCR amplifiable, and thus analyzable, nuDNA content at all. However, the advantages of the 454-based approaches come at a cost that is directly linked to DNA degradation. Specifically, in contrast to PCR-based approaches, which in theory can be successfully performed as long as a single, initially amplifiable template molecule is present in a DNA extract, 454-based analyses (in their current avatar as the Roche GS20 and FLX platforms) require a large amount of initial DNA (e.g., 1 µg) from which to construct the initial DNA libraries. While this template naturally need not all be copies of a single specific DNA sequence, the endogenous DNA content of ancient samples is predominantly much lower than the one that can be obtained from modern samples, due to both the facts that DNA degradation has occurred and that often the only source material available is bone (with a relatively lower DNA content than other tissues).

14.4

Degradation and Sequencing Accuracy

In addition to simply reducing the absolute size and quantity of template molecules available for analysis, DNA degradation plays another role that can have serious



effects on the data generated in paleogenomic studies. Since the earliest studies that investigated the qualities of DNA recovered from ancient remains [17, 18], it has been known that a small, but apparently common, number of processes can damage DNA molecules in such a way that the sequence is modified, yet it is still PCR amplifiable and thus sequenceable. Although these reactions do not hinder PCR, they do lead to sequence modification, as a result of which the generated sequence differs from the

undamaged original template. A number of recent studies have demonstrated that these differences originate predominantly from the hydrolytic deamination of cytosine to uracil or its analogues, although a number of other modifications may play a role [19–25]. Owing to the complementary nature of DNA, such deamination of cytosine to uracil is manifested in amplified sequences as cytosine-to-thymine and guanine-to-adenine transitions. To complicate matters further, all DNA polymerases have innate error rates, leading to a small number of nucleotide misincorporations during replication of the DNA molecule. Enzyme and template errors are rarely an issue in conventional PCR-based studies due to several reasons. First, with a few possible exceptions (cf. [26, 27]), the errors are essentially randomly distributed along the template molecules, and the number of DNA templates at the commencement of PCR reactions is normally sufficiently high, so that for every damaged nucleotide position, a large number of undamaged molecules exist. Therefore, the final sequence of the resultant amplicons will be generated from predominantly unmodified templates and will not show errors (Figure 14.2a). While exceptions do exist, including when PCR starts from very low numbers of template molecules (Figure 14.2b and c), or amplicons are cloned prior to sequencing, this is the general rule. In contrast, however, the very nature of emPCR renders the combined effect of DNA miscoding lesions and sequencing error important. Specifically, following library construction, the resultant emPCR in general represents PCR amplification from single template molecules. As all descendent molecules within a single emPCR lipid micelle, and as such, the resultant sequence of this micelle, derive from these single templates, any initial damage or sequencing error will be represented in the sequence, resulting in the generation of erroneous data (Figure 14.2d–f). Unfortunately, the problem is not this alone. Two recent papers have demonstrated that, in addition to DNA damage and enzyme error, a third mechanism exists that further adds to this problem. In particular, the enzymatic treatment required for library construction prior to emPCR requires the use of *Escherichia coli* DNA polymerase I Klenow fragment to fill nicks at the template–adapter 5′–3′ junction. Thus, as argued by Brotherton *et al.* [24] and Briggs *et al.* [25], not only does this effectively ensure

Figure 14.1 DNA degradation. (a) Calculated half-lives of nucleotides when degraded by deamination at different temperatures. The half-life ($T_{1/2}$) is calculated as $T_{1/2} = \ln(2)/k$, where k is the rate constant at any particular temperature. The relationship between k and temperature is described using the Arrhenius equation, $k = Ae^{-E_a/RT}$, where A is the Arrhenius constant, E_a is the activation energy, R the gas constant, and T the temperature (in K). Extrapolated from the data of Frederico *et al.* [46]. The figure demonstrates clearly that as the temperature of the reaction is decreased, the $T_{1/2}$ of nucleotides increases exponentially. Under the assumption that deamination is one of the key DNA

damaging processes to affect ancient samples, the data demonstrate how the sample storage temperature after death will significantly affect the survival time of DNA molecules. (b) The relationship between absolute starting template copy number of PCR amplifiable molecules in an ancient DNA extract, with size of amplicon. Measurements (black diamonds) of mtDNA content within a DNA extract were taken from a frozen Siberian mammoth bone [7] and fit to an exponential theoretical curve. The data demonstrate that the total number of PCR amplifiable DNA molecules in a sample increases exponentially with amplicon length.

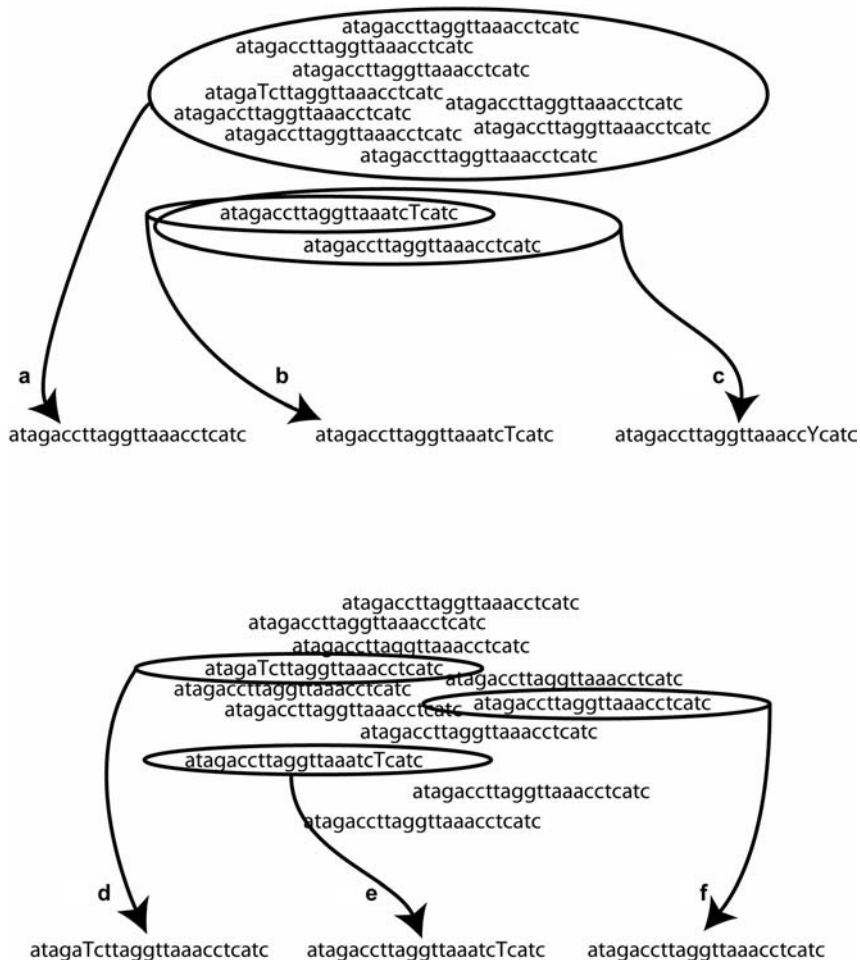


Figure 14.2 The effect of DNA damage-derived miscoding lesions on ancient DNA sequences. (a–c) During conventional PCR/direct Sanger sequencing approaches, damaged molecules do not affect the DNA sequence generated when they are rare in the original DNA template population, as the majority of molecules carry the undamaged base (a). However, when PCR starts from a single damaged molecule (b) or from a mixture of a small amount of damaged and undamaged molecules (c), the resultant

sequence will be either erroneous (b) or contain sequencing ambiguities (c). (d–f) During emPCR/454-based sequencing, each sequence that is generated derives from a single original template molecule. Therefore, any miscoding lesions present in the original molecules (e.g., d and e) will ultimately lead to the generation of erroneous sequences. These can be identified only when homologous undamaged sequences are also generated (e).

template–adapter ligation but should any single-stranded breaks or nicks with 3′-OH exist within the ancient template (extremely likely due to DNA degradation postmortem), these will also act as a Klenow fragment extension site leading to the generation of further sequencing errors. The details of the mechanism are beyond the scope of

this chapter, but seem extremely plausible. For further details, refer to the primary literature [24, 25].

14.5

Sample Contamination

The effect of DNA degradation on potential paleogenomic studies has one further consequence, DNA contamination. Within the framework of classical aDNA studies, contamination problems refer predominantly to a quantity of template molecules that exhibit sufficient similarity to the endogenous target molecules so as to be either preferentially amplified or coamplified during PCR, leading to the generation of misleading results (cf. [28] for more details). This problem has attracted a growing body of research investigating, among other things, the mechanics behind contamination and potential solutions to the problem (e.g., [29–34]). Such contamination is a significant headache for conventional studies, so much so that one of the pleasant benefits of conventional 454-based approaches is that the lack of predetermined target-specific primer-based amplification makes this problem nearly irrelevant. However, in contrast, contamination presents an altogether different problem for paleogenomic studies. In particular, the problem stems from the very untargeted nature of the method. Owing to the fact that in its unmodified format libraries are constructed from unmodified DNA extracts, the content of the DNA extract becomes extremely important to the analysis. With modern, high-quality samples, researchers can be confident that a large, if not complete, proportion of the DNA in their extract derives from the target. As such, the library will contain nearly 100% target DNA molecules and the resultant sequences will, for the most part, represent the target DNA (Figure 14.3a). Unfortunately though, this is a luxury that paleogenomic studies do not have. Almost without exception, ancient tissue samples contain a large amount of nonendogenous DNA. The sources and types vary, but intuitively can be divided into two broad classes. The first, and probably dominant, is sources of bacterial or other environmental organism DNA derived either from the environment the samples have been preserved in or potentially even from initial sample putrefaction [29]. Although tissues such as bone and tooth may appear solid to the untrained eye, it is often a surprise to find that their consistency is more akin to sponge, with degraded samples containing up to 50% air [35]. This not only acts as a store for the DNA of organisms that were present during putrefaction but also likely acts as a natural capillary system for drawing in further sources of DNA from the soil. The second potential source is DNA derived from human (or other) handling once excavated/sampled, or even reagents or conservation treatments that may contain DNA [36, 37]. Regardless of the source, the effect of this contamination is extensive and can be readily observed in the data of the initial Neandertal and mammoth paleogenomic publications (Figure 14.3b and c). For example, the mammoth bone chosen for the initial study by Poinar *et al.* [7] was recovered frozen from Siberian permafrost, had been kept frozen since recovery, and had been chosen from among a number of others specifically for its believed high-quality endogenous DNA content

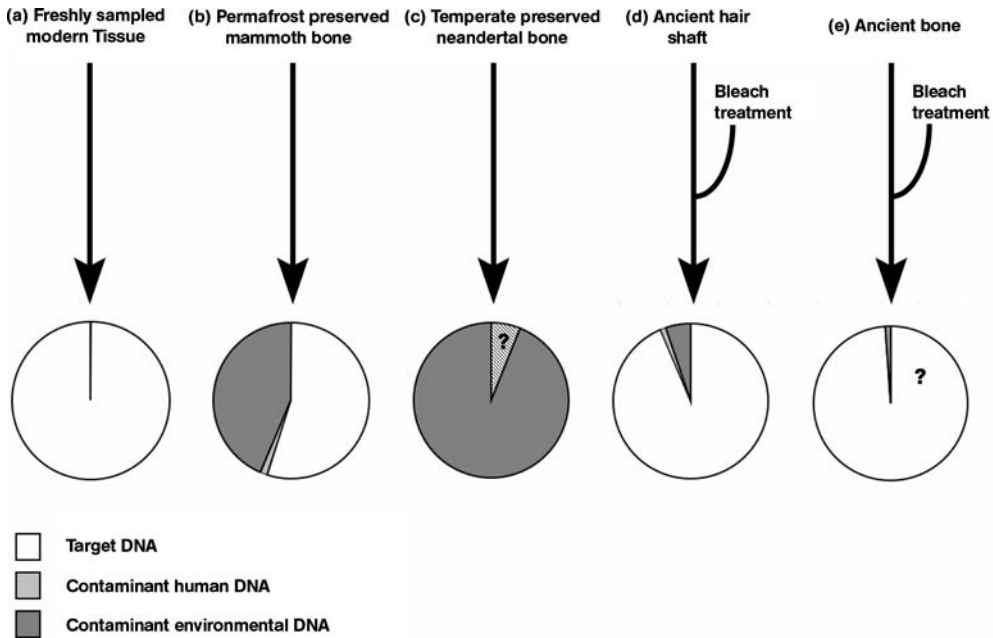


Figure 14.3 Observed, estimated, and hypothetical sequence distributions from different tissues. The distribution of sequences of different origin among 454-sequenced DNA extract varies considerably by DNA source. While freshly sampled modern tissues are expected to yield up to 100% target DNA sequences (a), in older bone samples the contribution of modern human and environmental DNA sources can be extremely high (b and c), from $\approx 45\%$ in well-preserved permafrost frozen bones [7] to more than 95% in temperate-preserved bones [8]. One

solution that has been demonstrated is the use of less porous DNA sources when available, such as keratin containing hair shafts (unpublished estimates based on Ref. [41]), that can be readily decontaminated from nontarget DNA sources using dilute bleach solution (d). A potential solution for ancient bone is the bleaching of the powdered bone prior to DNA extraction [31, 32] (e), a method that may remove up to 99% of the contaminant DNA sequences [33]. How well this applies to paleogenomic analyses remains to be tested.

(chosen using a series of quantitative real-time PCR prescreening analyses). Thus, this sample likely represents one of the highest quality ancient bone samples available for analysis. Regardless, nearly 55% of the sequences generated in the initial GS20 analysis of the library were likely of nonmammoth origin. Of these, 1.4% were alignable to the human mtDNA and nuDNA genome, representing either human contaminants derived from handling or the DNA and library preparation processes, or possibly mammoth nuDNA sequences too closely related to human nuDNA to be discriminated. The remainder were presumably derived either from organisms present during the sample's original putrefaction or from the environment in the time since (Figure 14.3b). With regard to "poorer" quality samples, the levels of contaminants are shockingly high – for example, only 6.2% of the 254 933 sequences generated from the temperate environment preserved, 38 000-year-old Neandertal bone DNA extract by Green *et al.* [8] aligned with primate DNA, of which

an unknown, although likely high [38], proportion derived from anatomically modern human (AMH) contaminant DNA (Figure 14.3c). Not only does such inefficiency have clear economical ramifications, vastly increasing the cost per target base pairs sequenced, but also does have implications for data quality. Consider the Neandertal examples. The close relationship of AMHs to Neandertals suggests that a large amount of their nuDNA is identical. Therefore, the presence of such high levels of contaminant sequences, much of which can be directly aligned to primates, leads to the implication that true Neandertal sequences can only be identified where nucleotide differences exist that discriminate them from AMH sequences and where these differences can be discounted as being derived from damage, sequencing error, or other causes of sequence modification. As such, the efficiency of the assays drops further still.

In summary, the key challenges facing paleogenomics are DNA survival, resultant sequence quality, contamination, and thus the economical viability of the studies. Fortunately, despite paleogenomics being at its infancy, a number of solutions to its problems have already been demonstrated, or potentially exist.

14.6

Solutions to DNA Damage

Of all the problems, the one that is hardest to directly solve is DNA damage. Although a number of repair systems exist both *in vivo* and *in vitro* that might in theory be applied to aDNA, they are not yet at the level where they have been successfully applied to ancient DNA templates in a way that can then be used in paleogenomic studies. This is predominantly because the simplest systems, for example, the enzymatic treatment of the DNA to repair nicks, breaks, or other PCR blocking damage [39], are currently at a level where, in repairing the DNA, they introduce further sequence errors (cf. [40]). Therefore, with regard to samples for use in analyses, the simplest solution for now is to choose well-preserved specimens that will not require repair. The fact that studies have been successful on both frozen and temperate environment preserved remains, including some that date back beyond the limits of ^{14}C dating (approximately 60 000 ^{14}C years [41]), is at least encouraging in this regard. With regard to the challenges presented by miscoding lesions, sequencing error, and other processes that directly modify the templates that can be used in paleogenomic analyses, several theoretical solutions exist. The first is the treatment of template molecules prior to the analysis with enzymes that cleave the templates where damage is present (e.g., the treatment of templates with the enzyme uracil-*N*-glycosylase or analogues to cleave uracils derived from cytosine deamination). However, the probable drawback of this is a significant reduction in the numbers of template molecules available for library construction and subsequent emPCR, due to the likely large-scale occurrence of uracils in old templates. An alternative solution derives from one of the great advantages in 454-based approaches, the very fact that a large number of independently sequenced molecules are produced during each analysis. Therefore, in theory, it should be possible to

obtain multiple sequence coverage for each sequence generated (Figure 14.4). Unfortunately, however, given the current state of paleogenomics, this is easier said than done. For example, the large size of the genomes of the organisms most likely to be under paleogenomic study in comparison to the data generated in an FLX run (3 gigabases compared to 100 megabases) means that even with 100% sequencing efficiency, the recovery of multiple covered nuDNA sequences will be slow and expensive (in terms of both time and financial costs). With lower efficiency, due predominantly to contamination, this problem is exasperated. One potential exception is organelle genomes. Organelle genomes are much smaller than nuclear genomes (e.g., approximately 16–17 kb in many mammals) but are present with a high number of copies in each cell. The total organelle genome content in a cell varies *in vivo* with tissue type and organism, and postmortem potential differences in degradation rates may affect this further. However, in theory, the organelle genome content in an extract should be reflected in the frequency with which organelle sequences are observed among the total number identified to the endogenous DNA. Several studies on mammoth DNA have shown the number to vary by tissue, from approximately 1 in every 600–700 mammoth sequences recovered for bone [7, 23] to a much higher 1 in every 50–100 for hair shaft [41]. These relatively high occurrences, combined with the small organelle genome sizes, enable multiple sequence coverage to be obtained fairly easily, especially in situations where contamination levels can be reduced (discussed later). Once multiple coverage is obtained for a DNA sequence of interest, it becomes trivial to discriminate true from artificial sequences, through a simple consensus analysis (Figure 14.4).

14.7

Solutions to Contamination

The problem of contamination has been approached in several ways, with several further potential solutions remaining to be tested. Unsurprisingly, due to the challenge of discriminating AMH and Neandertal DNA sequences from each other, the pioneering studies on Neandertal nuDNA [8, 9] focused specifically on this problem and used similar arguments to help discriminate the true Neandertal nuDNA from modern human contaminants. For example, Green *et al.* [8] took a combined approach that used amino acid racemization plus conventional mtDNA PCR and sequence analysis of the potential Neandertal bones. Arguments about the efficacy of racemization as a tool for estimating aDNA quality notwithstanding [42], the authors combined racemization indicators of DNA survival with results of PCR analysis using primers for both human and Neandertal mtDNA that amplify 63 and 119 bp fragments. By analyzing cloned sequences of the amplicons, they identified a Neandertal bone in which 99% of the short mtDNA fragments, and 94% of the long fragments, contained Neandertal specific sequences. As such, the authors then inferred from this that AMH mtDNA, and thus presumably nuDNA, was extremely rare in the sample. Using a similar PCR-based analysis, Noonan *et al.* [9] concluded similarly that only 2% of the sequences were of modern human origin.

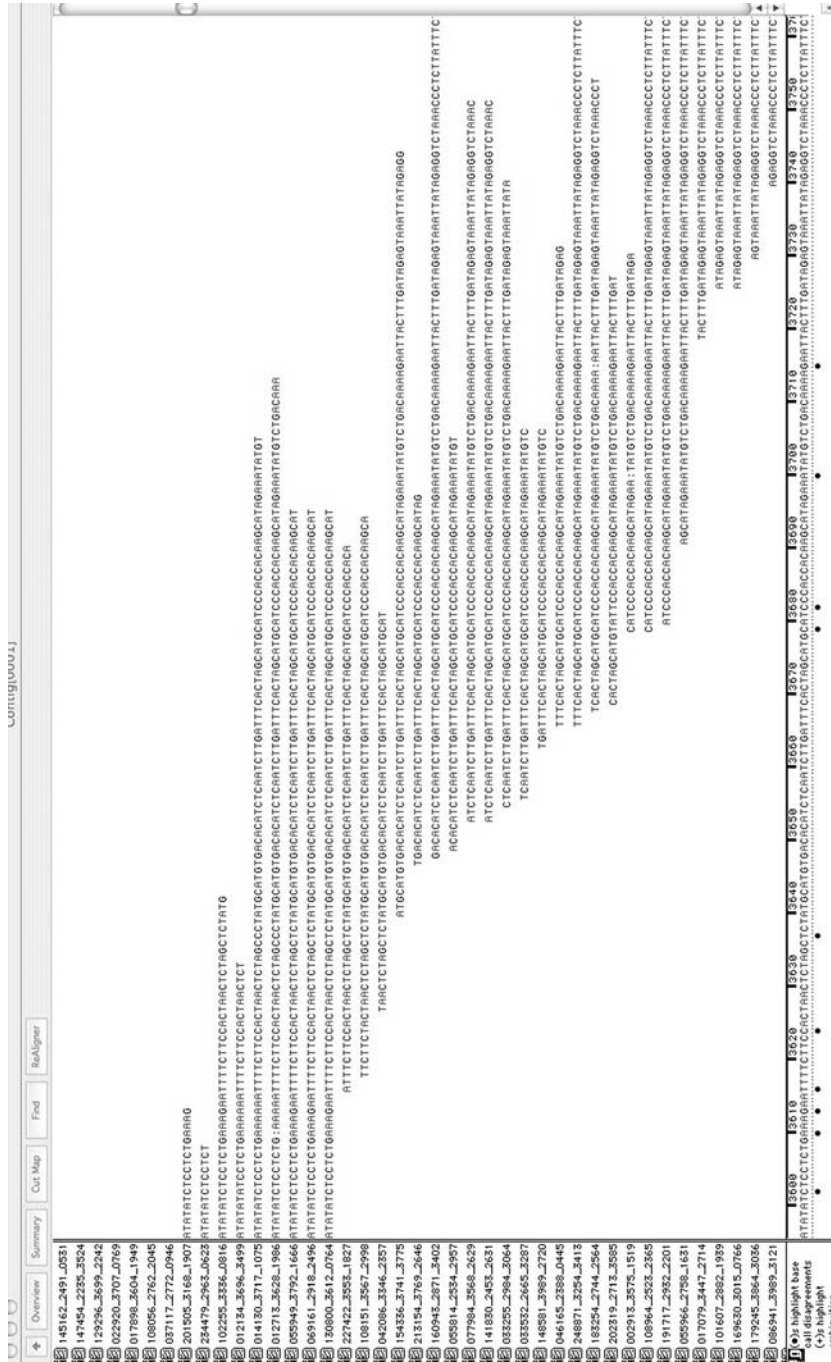


Figure 14.4 Multiple sequence coverage enables artifacts in the sequence to be easily identified. An alignment of independently generated 454 sequences that map to the mammoth mtDNA genome is shown. At each position of the mtDNA genome, sequence coverage is sufficient that artifacts can be easily identified among the consensus “correct” nucleotides. Data taken from Ref. [41] and aligned using SEQUENCHER [47].

Despite these analyses, however, the efficacy of the methods used has been questioned, in part as a result of the different conclusions drawn by the two studies with regard to questions such as the divergence times of Neandertals from AMHs and the contribution of the Neandertal genome into the modern gene pool. Specifically, while the data of Noonan *et al.* [9] enabled an estimation of an AMH–Neandertal divergence time of 706 kya, and 0% contribution of Neandertals to the modern European AMH gene pool, the data of Green *et al.* [8] can be used to argue for a younger and more substantial contribution (516 kya, and with Neandertal nuDNA sequences carrying the derived allele for approximately 30% of human SNPs in the public domain). One recent study [38] has argued that this discrepancy is derived from higher than first estimated levels of modern human contaminant DNA in the data sets, specifically that of Green *et al.* [8]; for example, the critics argue that when the Green *et al.* [8] Neandertal sequences are compared with the HapMap SNP database, they have the derived allele approximately 33% of time, while the human reference sequence has the derived allele 38% of time. In contrast, the comparable percentage in the Noonan *et al.* [9] data is only approximately 3%. In conclusion, the authors argue that the Green *et al.* [8] data suffer to a large extent from contamination and other forms of sequencing error.

In the light of the apparent problems with the above solutions, do others exist? Fortunately, the answer is yes. Several alternative solutions are possible. In a recent study, it has been demonstrated that, when possible, the choice of tissue used in the analysis can play a key role in reducing the contaminant load. On the basis of the observations from previous forensic genetic and ancient DNA studies [43–45], Gilbert *et al.* [41] hypothesized that the adoption of hair shaft material as a source of paleogenomic DNA would have several key advantages over bone material, including enriched mtDNA content, reduced DNA damage levels, and, more relevant, the property of being easily decontaminated from environmental DNA. Using a data set of 10 permafrost mammoth hair samples, the authors demonstrated that following the adoption of a simple hair wash in dilute bleach (10–20 × dilution of commercial strength) prior to DNA extraction, extremely pure mammoth DNA could be recovered from the samples (Figure 14.3d). As such, this enables the rapid accumulation of both well-supported mtDNA complete genomes and nuDNA, setting perhaps for the first time the stage for relatively efficient ancient nuDNA genome sequencing. The arguments that the authors used to explain these findings center around particular biological and chemical properties of keratinous tissues, and as such, it is likely that in addition to hair, other tissues that may exhibit similar properties are hoof, horn, feather, beak, and nail. Incidentally, it is worth mentioning that from a sampling point of view, hair sampling at least is less destructive than bone sampling, which is much harder to detect visually.

Naturally, such tissues are of use only when they exist, but when the choice is available they appear to be superior to at least bone or tooth. But what if bone is the sole choice? Several potential options exist. For instance, Noonan *et al.* [9] have

demonstrated the option of using target-specific oligonucleotides to “fish” out specific target sequences of interest from the DNA extract (in this study, once PCR amplified from metagenomic libraries, but the technique might equally be applicable to the original aDNA extract). In this way, sequences of interest (if known *a priori*) can be enriched away from the general contaminants and following suitable amplification steps can be used in high-throughput sequencing. An alternative, as yet untested, method is the adoption of conventional aDNA protocols that involve the bleach treatment of bone powder prior to DNA extraction [31, 32] (Figure 14.3e). Based on the concept that the endogenous DNA is largely protected from the bleach within crystal aggregates or other calcified structures, while the exposed contaminants are degraded, this method is effective, although at a cost. Using conventional quantitative real-time PCR-based analysis of known contaminants versus endogenous DNA, one study has demonstrated that although 99% of contaminant sequences are degraded following treatment, so are 75% of host DNA sequences [33]. Quite how well the efficacy of the method translates to 454 analysis remains to be seen, but the method does appear to be an option, if enough starting material can be obtained to offset the loss of endogenous DNA.

14.8

What Groundwork Remains, and What Does the Future Hold?

Regardless of the challenges that have been faced by the initial paleogenomic studies, and potential questions that hang over the empirical conclusions of the research, the initial studies have been groundbreaking in their vision and have demonstrated that an exciting future faces the field of ancient DNA. Clearly, a number of basic groundwork studies require the attention of the field, specifically with regard to resolving the problem of background contamination in ancient bone and teeth samples. Advances in sample decontamination techniques, or the development of methods to specifically isolate the true endogenous DNA from the contaminant molecules, will greatly increase the efficiency of future studies. In addition to the bleach methods already suggested above, under the (possibly unsupported) assumption that the true sequences are more degraded than the contaminants, thus shorter in length, other options include the use of bead- or filter-based methods to physically isolate the fractions based on size alone. Of similar use will be further study into the use of alternative materials as a DNA source. A third avenue worth exploring, which will benefit both paleogenomics and conventional use, is the development of methods to ensure that smaller initial template is required for the construction of successful DNA libraries prior to emPCR amplification. Undoubtedly, as the technology becomes more widely used and developed, this will happen. Furthermore, it is of course to be hoped that in concordance, the price of the method (not mentioned, but a clear hindrance to paleogenomic study!) will fall, opening up its use to a wider audience. Finally, and of great interest, is how paleogenomics will perform

in the face of current competing, and as yet undeveloped, sequencing technologies. With their specific application on smaller DNA templates in comparison to the GS20 and FLX platforms, the Solexa and SOLiD platforms will be of particular interest as long as sufficient bioinformatics tools are developed to help parse and assemble the already problematic sequences that will comprise the generated data.

In the light of the developments so far, and those to come, the future of paleogenomics is bright. It is not inconceivable that the next few years will start to see the publication of the first ancient and extinct nuDNA genomes, enabling for the first time the recovery of lost genetic information on this unprecedented scale. At the same time, aDNA studies have been placed in the position where they can move toward complete mtDNA or other organelle genome analysis of ancient species and populations, greatly improving the analytical power that can be applied to evolutionary or other questions.

References

- Noonan, J., Hofreiter, M., Smith, D., Priest, J.R., Rohlan, N., Rabeder, G., Krause, J., Dettler, J.C. *et al.* (2005) Genomic sequencing of Pleistocene cave bears. *Science*, **309**, 597–599.
- Haddrath, O. and Baker, A.J. (2001) Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **268**, 939–945.
- Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J. and Ward, R. (2001) Complete mitochondrial genome of two extinct moas clarify ratite evolution. *Nature*, **409**, 704–707.
- Krause, J., Dear, P., Pollack, J.L., Slatkin, M., Spriggs, H., Barnes, I., Lister, A.M., Ebersberger, I. *et al.* (2006) Multiplex amplification of the mammoth mitochondrial genome and the evolution of the Elephantidae. *Nature*, **439**, 724–727.
- Rogaev, E.I., Moliaka, Y.K., Malyarchuk, B.A., Kondrashov, F.A., Derenko, M.V., Chumakov, I. and Grigorenko, A.P. (2006) Complete mitochondrial genome and phylogeny of Pleistocene mammoth *Mammuthus primigenius*. *PLoS Biology*, **4**, e73.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., Macphree, R.D.E., Buiges, B., Tikhonov, A., Huson, D.H. *et al.* (2006) Metagenomics to palaeogenomics: large-scale sequencing of mammoth DNA. *Science*, **311**, 392–394.
- Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M. *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature*, **444**, 330–336.
- Noonan, J., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D. *et al.* (2005) Sequencing and analysis of Neanderthal genomic DNA. *Science*, **314**, 1113–1118.
- Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature*, **362**, 709–715.
- Smith, C., Chamberlain, A.T., Riley, M.S., Cooper, A., Stringer, C.B. and Collins, M.J. (2001) Neanderthal DNA: not just old but old and cold? *Nature*, **410**, 772–773.

- 12 Smith, C., Chamberlain, A.T., Riley, M.S., Stringer, C. and Collins, M.J. (2003) The thermal history of human fossils and the likelihood of successful DNA amplification. *Journal of Human Evolution*, **45**, 203–217.
- 13 Cano, R.J., Poinar, H. and Poinar, G.O. Jr. (1992) Isolation and partial characterization of DNA from the bee *Problebeia dominicana* (Apidae: Hymenoptera) in 25–40 million year old amber. *Medical Science Research*, **20**, 249–251.
- 14 DeSalle, R., Gatesy, J., Wheeler, W. and Grimaldi, D. (1992) DNA sequences from a fossil termite in Oligo-Miocene amber and their phylogenetic implications. *Science*, **257**, 1933–1936.
- 15 Woodward, S.R., Weyand, N.J. and Bunnell, M. (1994) DNA sequence from Cretaceous period bone fragments. *Science*, **266**, 1229–1232.
- 16 Kim, S., Soltis, D.E., Soltis, P.S. and Suh, Y. (2004) DNA sequences from Miocene fossils: an *ndhF* sequence of *Magnolia latahensis* (Magnoliaceae) and an *rbcL* sequence of *Persea pseudocarolinensis* (Lauraceae). *American Journal of Botany*, **91**, 615–620.
- 17 Pääbo, S. (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 1939–1943.
- 18 Pääbo, S., Higuchi, R. and Wilson, A. (1989) Ancient DNA and the polymerase chain reaction. *The Journal of Biological Chemistry*, **264**, 9709–9712.
- 19 Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. and Pääbo, S. (2001) DNA sequences from multiple amplifications reveal artefacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research*, **29**, 4793–4799.
- 20 Gilbert, M.T.P., Hansen, A.J., Willerslev, E., Rudbeck, L., Barnes, I., Lynnerup, N. and Cooper, A. (2003) Characterisation of genetic miscoding lesions caused by post mortem damage. *American Journal of Human Genetics*, **72**, 48–61.
- 21 Binladen, J., Wiuf, C., Gilbert, M.T.P., Bunce, M., Barnett, R., Larson, G., Greenwood, A.D., Haile, J. *et al.* (2006) Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. *Genetics*, **172**, 733–741.
- 22 Stiller, M., Green, R.E., Ronan, M., Simons, J.F., Du, L., He, W., Egholm, M., Rothberg, J.M. *et al.* (2006) Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 13578–13584.
- 23 Gilbert, M.T.P., Binladen, J., Miller, C., Wiuf, C., Willerslev, E., Poinar, H., Carlson, J.E., Leebens-Mack, J.H. *et al.* (2007) Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. *Nucleic Acids Research*, **35**, 1–10.
- 24 Brotherton, P., Endicott, P., Sanchez, J.J., Beaumont, M., Barnett, R., Austin, J. and Cooper, A. (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Research*, **35**, 5717–5728.
- 25 Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J. *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 14616–14621.
- 26 Gilbert, M.T.P., Willerslev, E., Hansen, A.J., Rudbeck, L., Barnes, I., Lynnerup, N. and Cooper, A. (2003) Distribution patterns of postmortem damage in human mitochondrial DNA. *American Journal of Human Genetics*, **72**, 32–47.
- 27 Gilbert, M.T.P., Shapiro, B., Drummond, A. and Cooper, A. (2005) Post mortem DNA damage hotspots in Bison (*Bison bison* and *B. bonasus*) provide supporting evidence for mutational hotspots in

- human mitochondria. *Journal of Archaeological Science*, **32**, 1053–1060.
- 28 Hofreiter, M., Serre, D., Poinar, H.N., Kuch, M. and Pääbo, S. (2001) Ancient DNA. *Nature Reviews. Genetics*, **2**, 353–358.
 - 29 Gilbert, M.T.P., Rudbeck, L., Willerslev, E., Hansen, A.J., Smith, C., Penkman, K.E.H., Prangenberg, K., Nielsen-Marsh, C.M. *et al.* (2005) Biochemical and physical correlates of DNA contamination in archaeological human bones and teeth excavated at Matera, Italy. *Journal of Archaeological Science*, **32**, 783–795.
 - 30 Gilbert, M.T.P., Hansen, A.J., Willerslev, E., Turner-Walker, G. and Collins, M. (2006) Insights into the processes behind the contamination of degraded human teeth and bone samples with exogenous sources of DNA. *International Journal of Osteoarchaeology*, **16**, 156–164.
 - 31 Salamon, M., Tuross, N., Arensburg, B. and Weiner, S. (2005) Relatively well preserved DNA is present in the crystal aggregates of fossil bones. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 13783–13788.
 - 32 Malmström, H., Stora, J., Dalén, L., Holmlund, G. and Götherström, A. (2005) Extensive human DNA contamination in extracts from ancient dog bones and teeth. *Molecular Biology and Evolution*, **22**, 2040–2047.
 - 33 Malmström, H., Svensson, E., Holmlund, G., Gilbert, M.T.P. and Götherström, A. (2007) More on contamination: asymmetric molecular behaviour as a mean to identify authentic ancient human DNA. *Molecular Biology and Evolution*, **24**, 998–1004.
 - 34 Kemp, B. and Glenn Smith, D. (2005) Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. *Forensic Science International*, **154**, 53–61.
 - 35 Turner-Walker, G., Nielsen-Marsh, C.M., Syversen, U., Kars, H. and Collins, M.J. (2002) Sub-micron spongiform porosity is the major ultra-structural alteration occurring in archaeological bone. *International Journal of Osteoarchaeology*, **12**, 407–414.
 - 36 Leonard, J.A., Shanks, O., Hofreiter, M., Kreuz, E., Hodges, L., Ream, W., Wayne, R.K. and Fleischer, R.C. (2007) Animal DNA in PCR reagents plagues ancient DNA research. *Journal of Archaeological Science*, **34**, 1361–1366.
 - 37 Nicholson, G.J., Tomiuk, J., Czarnetzki, A., Bachmann, L. and Pusch, C.M. (2002) Detection of bone glue treatment as a major source of contamination in ancient DNA analyses. *American Journal of Physical Anthropology*, **118**, 117–120.
 - 38 Wall, J.D. and Kim, S.K. (2007) Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genetics*, **3**, e175.
 - 39 D'Abbadie, M., Hofreiter, M., Vaisman, A., Loakes, A., Gasparutto, D., Cadet, J., Woodgate, R., Pääbo, S. *et al.* (2007) Molecular breeding of polymerases for amplification of ancient DNA. *Nature Biotechnology*, **25**, 939–943.
 - 40 Gilbert, M.T.P. and Willerslev, E. (2007) Rescuing ancient DNA. *Nature Biotechnology*, **25**, 872–874.
 - 41 Gilbert, M.T.P., Tomsho, L.P., Rendulic, S., Packard, M., Drautz, D.I., Sher, A., Tikhonov, A., Dalén, L. *et al.* (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science*, **317**, 1927–1930.
 - 42 Collins, M.J., Waite, E.R. and van Duin, A.C.T. (1999) Predicting protein decomposition: the case of aspartic-acid racemization kinetics. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **354**, 51–64.
 - 43 Wilson, M.R., Polanskey, D., Butler, J., DiZinno, J.A., Replogle, J. and Budowle, B. (1995) Extraction, PCR amplification and sequencing of mitochondrial DNA from human hair shafts. *BioTechniques*, **18**, 662–669.
 - 44 Jehaes, E., Gilissen, A., Cassiman, J.J. and Decorte, R. (1998) Evaluation of a decontamination protocol for hair shafts before mtDNA sequencing. *Forensic Science International*, **94**, 65–71.

- 45 Gilbert, M.T.P., Menez, L., Janaway, R.C., Tobin, D.J., Cooper, A. and Wilson, A.S. (2006) Resistance of degraded hair shafts to contaminant DNA. *Forensic Science International*, **156**, 208–212.
- 46 Frederico, L.A. Kunkel, T.A. and Shaw, B.R. (1990) A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*, **29**, 2532–2537.
- 47 SEQUENCHER v4.7 Build 2947 (2006) Gene Codes Corporation, Ann Arbor, MI, USA.

15

ChIP-seq: Mapping of Protein–DNA Interactions

Anthony Peter Fejes and Steven J.M. Jones

15.1

Introduction

Prior to the sequencing of the human genome [1], it was commonly assumed that the genetic code of a species contained all of the information required for the development of a single organism. On the surface, this assumption was reasonable: all cellular components are derived from the instructions contained within the genome. However, it has become increasingly clear that the control of the genome itself is managed through much more complex interactions, involving modifications to the DNA and its associated proteins [2], and that protein–DNA interactions have huge impacts on the phenotype of an individual [3, 4]. These modifications have the ability to regulate gene expression, turning genes on and off in ways that would be impossible to predict from the genome sequence alone. Compounding the complexity of the problem are the interactions of the machinery involved in gene expression. This includes transcription factors, which selectively interact with specific promoter, repressor, and enhancer DNA sequences to carry out their respective functions. Both epigenetics and gene regulation are now being studied by similar techniques, as they have the same fundamental goals: understanding and characterizing the nature of protein–DNA interactions involved in the regulation and mechanism of gene expression [5]. The method best suited for accomplishing these goals is the chromatin immunoprecipitation and massively parallel sequencing technique, known as ChIP-seq.

In this chapter, we will cover the history of the technique, and the general method used, followed by discussions of alternative protocols available, as well as some key works that have emerged in this field. Finally, we will explore some of the applications where ChIP-seq is likely to be used.

15.2

History

The ChIP technique generally uses a cross-linking agent to hold together DNA and the proteins with which it is in close contact, to allow the collection of specific proteins of interest by immunoprecipitation. The earliest targets of the ChIP protocol were not the DNA fragments pulled down by immunoprecipitation but the histone proteins. These proteins are now known to come together in specific combinations and undergo modifications, such as methylation, acetylation, and phosphorylation, which enable histones to participate both in a structural role and in more dynamic processes such as chromosome condensation during mitosis [6–8]. While the original aim of the ChIP protocol was the study of the histone structure and assembly, early ChIP users quickly discovered the persistent contact between histones and DNA when they were found not to dissociate by salt treatment [9]. Although the technology for investigating the immunoprecipitated DNA sequences was not available to early ChIP users, later researchers were able to carry out work on *cis*-acting transcriptional elements that began to provide insight into the control of transcriptional elements, such as promoters, enhancers, and repressors [10], and the role of chromatin regulation in gene expression [11].

The potential of ChIP to investigate DNA fragments was not overlooked. As new methods became available to study the immunoprecipitated DNA, ChIP was paired with methods such as Southern blotting, PCR, qualitative PCR, and, more recently, gene arrays [12, 13]. As each new technique for investigating mixed populations of DNA became available, interest in the chromatin immunoprecipitation increased. By the end of the 1990s, ChIP again became a popular technique to study proteins and both protein-associated DNA and RNA [14].

Most recently, ChIP has been combined with the emerging massively parallel sequencing techniques, called ChIP-seq. This protocol combines the ability to target specific proteins and their specific interacting DNA fragments with the high-throughput tag sequencing abilities of the next-generation sequencing machines to create an efficient and cost-effective whole genome view of DNA-binding sites for proteins of interest.

15.3

ChIP-seq Method

Chromatin immunoprecipitation is a relatively simple five-step process with the ability to collect fragments of DNA, known to be bound to a specific protein of interest, from living cells [14]. The first step is to treat the target cells or tissue with a cross-linking agent, typically formaldehyde, to form short covalent bonds between proteins or DNA molecules that are in contact. This selectively ties together compounds that are interacting *in vivo*, effectively freezing their interaction in time, ideal for capturing the short-lived transcription factor binding interactions. This step is followed by sonication or vortexing to disrupt the cells and physically shear

DNA to a desired size. The third step is the application of an antibody with selective affinity for the proteins of interest, to pull them down along with any other cellular entities to which they are attached. The antibodies are then collected, allowing nonantibody-bound proteins and loose DNA to be washed away. Exposure to high concentrations of salt and heat can be used to reverse the formaldehyde cross-linking, simultaneously disrupting the antibody binding. This results in a highly enriched collection of desired proteins and the DNA sequences to which they were bound *in vivo*.

When ChIP is paired with massively parallel sequencing, the DNA fraction is isolated and collected from the mixture, either by kit or phenol–chloroform extraction. The short DNA fragments will have an average length depending upon the length and vigor of the sonication or vortexing used during the shearing of the cells. The desired range of fragments can then be selected by running the DNA on an agarose gel and excising the portion containing the desired size, which is often in the range of about 150–500 bp.

An interesting consequence of the cross-linking is the detection of secondary protein–protein and protein–DNA interactions. Proteins in close proximity to the transcription factor of interest will become cross-linked during formaldehyde cross-linking and consequently will be pulled down along with the transcription factor of interest during the ChIP process. Any DNA to which they were also cross-linked will then be sequenced during the analysis of the immunoprecipitated DNA fragments. This creates multiple side-by-side points of enrichment of aligned reads along the genome, indicating the presence of the second DNA-interacting protein. Further analysis of these locations is likely to reveal that the binding motif of the targeted transcription factor is not present, indicating the DNA–protein interaction may not be direct.

An alternative method exists for the isolation of histones and their associated DNA, using a ChIP method that does not require a cross-linking step. In this protocol, described for postmortem brain tissue, the sonication or vortexing is replaced with a micrococcal nuclease digest, which does not disturb the coiling of nucleic acids on the histones of interest [15]. This allows the DNA to remain tightly wound around the histones throughout the entire ChIP process. After precipitation, the DNA can be separated from the histones by proteinase K treatment and extracted with phenol–chloroform. Although not applicable to transcription factor precipitations, this process can greatly simplify the collection of histones without disturbing their signaling modifications. Histone–DNA interactions investigated with this method also indicate that both the histone–DNA interactions and the histone modifications are relatively stable, and can be processed and successfully interpreted up to 30 h after the death of the donor.

15.4

Sanger Dideoxy-Based Tag Sequencing

One of the main reasons for the slow adoption of ChIP-based methods for analyzing DNA or RNA has been the inability to characterize the large populations of nucleic

acid fragments obtained [14]. Using Sanger dideoxy-based capillary sequencing to quantitatively analyze the millions of DNA fragments from a single ChIP experiment would be a costly process, complicated by the diversity in the population of fragments represented in the population. This is especially true for larger mammalian genomes. However, several methods have been developed to efficiently sequence and sample the fragments of interest. Although sequencing every fragment is rarely possible using dideoxy sequencing, it is certainly possible to sample them to gain an understanding of the composition of the collection.

One example of an early pioneering method for achieving this was used in the investigation of yeast telomeric heterochromatin, in which the precipitated DNA and a section of known yeast telomere were used as primer pairs to amplify and sequence regions near chromosome ends [16, 17]. This innovative method allowed the selective investigation of histone modifications near yeast chromosome telomeres, which would have been difficult, or impossible, with other techniques.

Unlike the above example, much of the work combining ChIP and Sanger dideoxy sequencing strategies is aimed at studying areas distant from the telomeric regions of chromosomes, and thus many of the strategies that have been developed include the creation of libraries containing cloned DNA fragments [18, 19]. Libraries have the clear advantages of being able to separate and amplify the many fragments from the original mixture in which they were isolated, as well as having flanking sequences for each cloned fragment that can be used as primers for sequencing reactions. However, the main drawbacks of Sanger-based investigations of DNA fragment libraries have been the large number of sequencing reactions required to obtain a statistically representative set of sequences from the ChIP-derived fragments and the significant time and effort required to create the library before sequencing can begin [20].

To avoid these limitations, more high-throughput methods for sequencing the library of fragments have also been developed. These methods forgo a portion of the sequencing length to gain an increase in sampling quantity. This is done by shortening each fragment to create tags, which can then be serially ligated before sequencing. In this way, a single sequencing reaction can detect the presence of a number of fragments by identifying the points of origin of each unique tag sequenced. This method of serially ligated tags is similar to that of the well-established Serial Analysis of Gene Expression (SAGE) technique [21, 22]. Three methods that combine ChIP and serially ligated tags include ChIP-SAGE [23], Serial Analysis of Binding Elements (SABE) [24], and ChIP-PET (paired end tag) [25, 26]. All three techniques take advantage of the tag-based concept of SAGE by including an extra cloning step in which short tags are extracted from each fragment through type II restriction digest to create small tags from the immunoprecipitated fragments, which can then be ligated and sequenced using a single dideoxy sequencing reaction to obtain a greater sampling with fewer sequencing reactions. The first method, ChIP-SAGE (also known as GMAT), uses a straightforward technique of collecting tags, but suffers from a loss in the resolving power that could be obtained from sequencing the full fragment: the site of the interaction can be narrowed down only to 500–1000 bp, as the site of the protein–DNA interaction can be up to 1000 bp from the end of the fragment from which the tag was created [27]. Similarly, the SABE uses an elegant SAGE-like method,

but introduces a further enrichment step in which only the immunoprecipitated fragments are attached to one of a pair of linkers and then are subtractively hybridized against the nonenriched DNA fragments. Only those pairs of fragments that are able to hybridize with complementary primers (i.e., both from the enriched, primer-treated fraction) will be amplified, eliminating any sequence that hybridizes with DNA sequences from the nonenriched pool of fragments. By including a restriction enzyme site in only one of the primers, the amplified fragments can be cleaved to create 18 bp monotags that can then be spliced together, creating ditags, which can be further concatenated to create SAGE-like strings of ditags. Typically, up to 30 ditags can be sequenced in a single concatamer, dramatically increasing the information available per sequencing reaction. Because these tags come from either end of the fragment in which the protein–DNA interaction occurred, it is possible to map the protein–DNA interaction to the genome more accurately by bracketing the interaction site between the observed tags. Similarly, the ChIP-PET method obtains information from both ends of the fragment, excising all but ~20 bp on either end. Unlike SABE, this method is accomplished by creating a library of vectors containing the precipitated fragments. Whether using SABE or ChIP-PET, each set of paired end tags allows improved mapping of the fragment’s origin back to the host genome, an automated task for which software exists [28].

15.5

Hybridization-Based Tag Sequencing

Until recently, the only alternative to Sanger-based sequencing was the use of DNA hybridization-based techniques, which depend on the base pairing or hybridization of a single-stranded DNA (or RNA) molecule with a complementary (or nearly complementary) DNA sequence to form a helical structure. The number of mismatches tolerated is referred to as the stringency of the hybridization, and can be manipulated to achieve the desired range of complementarity. This simple technique has been used for a wide variety of purposes, including sequencing applications in the kilobase range [29]. The simplest use of hybridization is the use of a single DNA molecule of known sequence, referred to as a probe, to locate the presence or anchor complementary sequences from among a mixture of fragments.

The early hybridization-based methods often used a single probe, and were applied to identify the presence or absence of a given sequence in a mixed sample [30, 31]. These techniques were not combined with ChIP, likely because of the limited utility of identifying individual sequence fragments before the completion of the human genome and the development of the faster, more comprehensive, hybridization-based gene arrays in the 1990s. The advent of the array technologies led to experiments in which a limited number of genes or sequences of interest were probed simultaneously, within the limit of available arrays [32]. While early gene arrays, also known as gene chips, were able to provide probes only for a small number of sequences, more modern arrays include nearly 2 million different probes. However, because of the combinatorial explosion in the number of possible sequences that arise as the length of the probes

increases, it is impossible to cover all possible genomic sequences in one array for many organisms. In addition, because of the necessity of including probes that are sufficiently long to allow unique identification of their origin, each gene array must carefully select which probes are included.

One recent example of the size to which arrays have grown is the Affymetrix Genome-Wide Human SNP Array 6.0, which has 1.8 million probes, including 906 600 probes that target known SNPs. Even the use of arrays of this magnitude for the analysis of ChIP-derived fragments leaves open two significant but related issues: (i) it is only possible to test the presence of sequences for which there is a probe and (ii) the limited number of probes that can be tested prevents truly fine-grained sequence search scans from being possible. It is also important to note that the resolving power of hybridization techniques depends on the size of the DNA fragments that are being probed: the larger the fragment the more likely it is to find a probe to which it will hybridize, but the poorer the resolution of data obtained.

Finally, users of hybridization-based techniques must be aware of difficulties experienced with the microarray platform itself. A common problem is the introduction of PCR-based biases, which can be difficult to quantify, caused by amplifying the DNA fragments before they are applied to the microarray [33]. Other difficulties have been reported in the detection of low-affinity or low-abundance tags [34, 35], and repeatability is often poor when processing the same sample with the same or different microarray [36, 37].

Despite these drawbacks, gene arrays have proven to be particularly useful when searching for known motifs or DNA sequences, such as in diagnostic applications or when working with small, fully sequenced genomes. In fact, DNA arrays have been used extensively with ChIP experiments performed in yeast and bacterial genomes, where it is possible to create DNA chips that contain all intergenic regions [13, 38]. This combination of chromatin immunoprecipitation and DNA chips is known as ChIP-chip or ChIP-on-chip. However, the use of DNA arrays is likely to be superseded by new sequencing methods based on the sequencing-by-synthesis approaches, which avoid many of the problems of hybridization approaches [39] while achieving resolutions estimated to be equivalent to a gene array containing ~1 billion probes [40].

15.6

Application of Sequencing by Synthesis

Sequencing by synthesis was first proposed and patented by Robert Melamede in 1985 [41], but it remained relatively unknown until the beginning of the millennium. Since then, it has undergone rapid development and commercialization, such that it is now possible to purchase a variety of sequencing devices that use variations of this basic concept. The most prominent companies offering the technology for sequencing by synthesis are 454 Life Sciences, acquired by Roche in 2007 [21, 42], and Solexa, acquired by Illumina in 2007 [43]. Each system offers significant advantages, and thus has found niche applications for which they provide excellent results. The 454 Life

Sciences machine (the Genome Sequencer FLX) is able to produce up to 400 000 reads per experiment, with an average length of 200–300 bp, which makes it ideal for *de novo* sequencing applications. In contrast, the Illumina 1G is able to produce up to 5 million reads per experiment and simultaneously performs eight separate experiments per sequencing run, with a maximum length of 50 bp. Most Illumina runs are performed with a length of 36 bp, ideal for applications in which only short sequencing reads are needed, as in the case of ChIP and other SAGE-like processes. Unique to the Illumina machine is the ability to divide a single run into eight separate experiments, as its reaction chamber is divided into eight lanes. This allows a single sequencing reaction to produce up to 5 million sequences per experiment or up to 40 million sequences per reaction.

This ability to sequence a very large number of DNA fragments in a massively parallel process is one of the major advantages of the nascent sequencing-by-synthesis method. In the context of a ChIP experiment, this enables researchers to obtain a saturating coverage of the immunoprecipitated DNA fragments very quickly. The low cost of performing massively parallel sequencing also makes ChIP-seq an effective strategy, compared to either hybridization or dideoxy-based sequencing methods, to achieve similar levels of coverage. It is also expected that sequencing-by-synthesis methods will continue to become less expensive, making ChIP-seq more accessible to researchers, while providing longer sequences and novel functionality such as paired end reads. These improvements will help provide an improved identification of the genomic source locations of fragments in repetitive regions, another feature that cannot be accomplished with a hybridization-based approach.

One of the first experiments combining ChIP with sequencing-by-synthesis technology was based on the modification of the combination of chromatin immunoprecipitation and paired end tag strategy, using 454-based sequencing [44] to map p53-binding sites in HCT1116 cells. However, unlike standard ChIP-PET and SAGE techniques in which multiple tag sets have to be ligated to form long concatamers, only two PETs were joined, forming a diPET. This drastically shortened the number of tags that could be identified in each sequencing reaction; however, the massively parallel nature of the sequencing reactions available on the 454 GS20 machine used in this experiment (between 200 000–300 000) more than compensated for the reduction in PETs sequenced per reaction. As a result, it was possible to identify 22 687 unique fragments, of which 8896 were uniquely mappable, identifying 57 clusters of sequenced reads that indicated likely p53-binding sites.

More recently, ChIP has been combined with the massively parallel sequencing-by-synthesis method used by the Illumina 1G machine. Unlike the ChIP-PET strategy, these Illumina-derived reads contain only a sequence from one end of each immunoprecipitated fragment. However, to offset this disadvantage, the Illumina 1G is able to generate a 10-fold or greater improvement in coverage, sequencing up to ~5 million reads per lane. This creates a much higher coverage for areas enriched through the ChIP process once sequenced reads are mapped back to their point of origin in the genome. Thus, instead of bracketing the site of a DNA–protein interaction between the two sequenced ends of a read, as would be observed in a PET-based sequencing approach, multiple fragments are found to overlap at the site of the interaction.

Because of the direct sequencing of single molecules in the enriched fragment by using the Illumina 1G, it is also possible to show that sequencing of each DNA fragment is directly analogous to counting the frequency at which a binding event is observed. Thus, the greater the number of fragments observed, the greater the number of times the event occurred at the time of the experiment.

Barski *et al.* [27] published the first demonstration of the ChIP-seq method, highlighting its speed and versatility, in May 2007. They were able to provide comprehensive genome-wide coverage data for more than 20 epigenetic marks, as well as the DNA-binding locations of the CTCF protein in human CD4+ T cells. Mainly focusing on methylation of the tail segment of histones, they were also able to demonstrate correlation between the presence of many of these marks with transcriptional activation or repression. This provides a clear indication of the role that histone modifications play in the regulation of gene activity. From their results, it is impossible to dismiss the relationship between histone methylation and the control of transcription, which marks a major step forward in understanding epigenetic signals in human cells.

A similar study was published by Johnson *et al.* [40] focusing on the neuron-restrictive silencer factor (NRSF, also known as repressor element-1 silencing transcription factor or REST). Using the ChIP-seq protocol and searching for areas of enrichment of the sequenced reads, they were able to map NRSF binding to 1946 locations in the Jurkat human T-lymphoblast cell line, with an estimated accuracy of ± 50 bp. One example of the power of this method was demonstrated by their ability to identify a degenerate binding site for a gene thought to be regulated by NRSF. Previous experiments were unable to identify any sequences that were likely to match NRSF-binding sites; however, an area of enrichment was observed upstream in the ChIP-seq experiment, showing that NRSF did indeed bind and regulate the gene as previously hypothesized. Indeed, 94% of predicted binding sites from the ChIP-seq method were found within 50 bp of an NRSF-binding motif, and virtually all sites with 90% or greater match to an NRSF motif were found to be occupied.

Robertson *et al.* [45] also published the results of their ChIP-seq experiment based on the STAT1 transcription factor. Their work on the STAT1 protein is similar to the work done with NRSF; however, STAT1 is regulated by phosphorylation, making the system more complex. Upon stimulation with interferon- γ , STAT1 residing in the cytoplasm is phosphorylated, causing it to migrate to the nucleus, where it gains the ability to form homodimers, heterodimers, and heterotrimers, which then bind the DNA. However, STAT1's phosphorylation is short lived and it is rapidly dephosphorylated, whereupon it dissociates from the DNA and returns to the cytoplasm.

This study was conducted on both interferon- γ stimulated and unstimulated Hela S3 cells, allowing a comparison between the two conditions. Sequenced reads found in the unstimulated sample were largely indistinguishable from noise, indicating that STAT1 binding is scarce without the interferon- γ stimulation; however, reads in the stimulated sample were observed to cluster into enriched areas, which passed false discovery rate thresholding. Thus, the ChIP-seq method is able to differentiate the two binding conditions and capture the short-lived binding during phosphorylation of the STAT1 transcription factor.

To facilitate the interpretation of the ChIP-seq data obtained in this experiment, Robertson *et al.* devised a method in which each short sequence read was assumed to be indicative of a fragment with a mean fragment length of 174 bp, determined experimentally. By extending each sequence read to a constant length, representative of the original fragment size used for Illumina sequencing, they were able to align sequenced reads back to the genome and look for areas of overlaps, which indicate enrichment. Using this process, binding sites appear as a Gaussian-like distribution, with median widths of about 40–50 bp and tails extending up to 1000 bp on either side, which compares favorably with typical ChIP-chip results of a single featureless peak of 500–1000 bp. The Gaussian-like distributions observed in this method can be used to locate a peak maximum, which was shown to coincide well with known locations of the STAT1 transcription factor, and the STAT1 binding motif predictions were generally found within 100 bp of the peak maxima. STAT1-enriched sites were also observed near known transcriptional start sites, with the highest density at –100 bp upstream, as would be expected for a transcription factor.

15.7

Medical Applications of ChIP-seq

Researchers are now able to observe genome-wide interactions between DNA and proteins *in vivo*, as well as changes in genetic regulation in response to various stimuli using ChIP-seq. It has begun to open the door for the development of genome-wide maps indicating histone modifications and locations of transcription factor, enhancer, repressor, and promoter-recruiting sequences. Undoubtedly, this knowledge will have a broad impact on our understanding of genomics-based medicine, leading to the development of novel treatments. There are already indications that our ever-improving understanding of the molecular mechanics of protein–DNA interactions is changing the medical field in diverse areas such as toxicology [46], aging [47], general health [48], and cancer [49, 50]. An increased awareness of the consequences of altering gene regulation will play a major role in the future study of each of these fields. In cancer, for example, our understanding of the genetic basis of the disease will require the study of key events such as the effect of transcription factor binding site mutations on oncogene regulation, the effect of *de novo* binding sites arising from mutations, and the epigenetic controls involved in oncogenesis. Clearly, the medical applications of the ChIP-seq method will have a transformative effect on how we perceive the study of human health.

15.8

Challenges

Many of the challenges for a broad application of the ChIP-seq are at the data alignment stage and the interpretation of data obtained from the next-generation sequencing devices. The most widely used tools for sequence alignment are based

on the venerable Smith–Waterman alignment algorithm, which provides an exact optimal solution, and the Basic Local Alignment Search Tool (BLAST), which provides a rapid near-optimal solution [51, 52]. Despite ongoing modifications to the BLAST algorithm [53, 54], the underlying strategy employed is not well suited to rapidly identify the genomic origin of short fragments. While comprehensive, the amount of time required to process millions of fragments quickly accumulates, making it unfeasible to utilize these methods for high-throughput sequence identification. Fortunately, a new generation of short-read alignment programs have been developed to accomplish this task.

The most popular of the alignment tools is ELAND (Efficient Local Alignment of Nucleotide Data) (Anthony J. Cox, Illumina Inc., 2007), which is able to rapidly identify the point of origin of a short sequence whenever a unique match with up to two base mismatches exists. When this condition is not met and there are equally likely points of origin for the fragment, no results are returned. For many data sets of mammalian DNA, this alignment tool allows between 50 and 65% of the fragments to be identified [40, 45]. Of the 24 million fragments observed by Robertson and colleagues while studying the STAT1 transcription factor, only 15 million were uniquely mapped back to the genome by the ELAND software. This is likely because ELAND is unable to align sequences containing insertions or deletions (gapped alignments), and the algorithm places a limit on the maximum number of mismatches (SNPs or sequencing errors) that may be accepted in any alignment. The ratio of mappable fragments to total sequenced fragments is expected to improve as new methods are developed for performing short-read alignment.

Indeed, there are already a plethora of competing packages, each with its own strengths including the Mosaik assembler (<http://bioinformatics.bc.edu/marthlab/Mosaik>), Exonerate [55], SXOligoSearch (Synamatix, Kuala Lumpur, Malaysia), and Slim Search (SLIM Search Ltd, Auckland, New Zealand). Many of these programs are already able to handle more complex functions, such as mapping a single fragment to multiple points of origin in the genome and the inclusion of insertions and deletions, and are thus beginning to overtake ELAND in use for short-read sequences.

A second major hurdle to the use of short-read sequences occurs during the interpretation of the data. Johnson *et al.* [40] devised a method in which a minimum threshold number of sequenced tags, set at 13, based upon ROC analysis [56] must be found within 100 bp. A second method, pioneered by Robertson *et al.*, uses the so-called “peaks,” in which observed sequences are assumed to extend to the mean fragment length of the precipitated DNA and the number of times each base position is observed in any extended read is collected into a histogram, from which the peaks can be identified. Peaks that have heights greater than a false discovery rate threshold are retained. Both methods permit a quick genome-wide scan of the template genome for areas of enrichment, representing the observed binding locations. An advantage of these methods is the simplicity in expressing these peaks or sequenced reads graphically, making them relatively simple to interpret visually. However, when reads that contribute to more than one nearby site overlap, or when secondary interactions exist, the peaks can become complex making it difficult to locate the true binding site of the protein of interest. Furthermore, as the sequencing depth

increases, fragments from nonspecific binding will begin to accumulate on the shoulders of each peak, making it difficult to find the peak region's boundaries. These issues remain to be addressed in the future.

A third major issue comes from the short nature of the sequences yielded by the Illumina process (currently 32–51 bp in length). These short reads often result in an insufficient amount of bases sequenced to be able to predict the true genomic point of origin of the DNA fragment. Related to this is the ambiguity inherent in determining the origin of DNA fragments that are derived from repeat regions in the genome. Similarly, it is also possible to sequence entire regions that do not exist in the template genome, or for which the sequence obtained contains sufficient mutations that the read shares too little identity with the analogous region in the template genome. Each of these situations causes a loss of information through the inability to interpret some subset of the results obtained from the interpretation. This is likely to be alleviated as the Illumina technology matures, or as competing technologies enter the market, enabling longer sequencing reads. Another possible solution is the use of a paired end read protocol during sequencing, which has the potential to provide additional fragments that may contribute to the identification of DNA fragments. At present, protocols for paired ends are under development for use with the Illumina 1G system, and will allow both ends of each fragment to be sequenced. In addition, the use of paired end data will likely provide an incremental improvement to the quality of alignments that are currently obtained and will assist in the disambiguation of fragment origins by ensuring that one of the two identified tags will be uniquely mappable. In the small number of cases where both tags are ambiguous, it may also be possible to find a unique site on the genome where the two tags are found in positions corresponding to the expected length of the fragment.

15.9

Future Uses of ChIP-seq

One challenge worth noting with respect to the ChIP-seq method relates to a further epigenetic change: the methylation of DNA. The methylation of cytosine residues is a well-known DNA modification that is thought to play a role in gene regulation. It is one of the best-studied epigenetic modifications of DNA across all organisms [57]. Although this experiment does not typically involve ChIP directly to pull down methylated bases, it is likely to become involved in the experimental protocol, nonetheless.

To study the methylation of DNA, the DNA of interest is treated with sodium bisulfite, which converts methylated cytosine residues to uracil [58]. Thus, the same segment of DNA can be sequenced twice, once with bisulfite treatment and once without, allowing the determination of which cytosines were converted to uracil and thus where the methylation existed along the DNA sequence. In the case of short-read sequencing technology, bisulfite-treated DNA is more difficult to interpret, as bisulfite causes a loss of information corresponding with the decrease in cytosines residues, and it is impossible to pair reads as would be done for Sanger

sequencing. Although it is difficult to map the slightly degenerate sequences back to the human genome, it is not impossible through exhaustive searches to identify likely sources of origin. However, reducing the amount of sequence space being searched using this method can be accomplished by performing a ChIP pulldown and treating only one-half of the sequences with bisulfite. The nonbisulfite-treated sequences can then be used to identify a small region of the genome the DNA is likely to have originated from, while the bisulfite-treated DNA can then be aligned against the reduced genomic regions to identify areas of cytosine methylation.

Similarly, with the advent of sequencing techniques that allow comprehensive enumeration of the DNA fragments, it is now possible to begin asking detailed questions about allelic frequency and SNPs within protein–DNA interaction sites [59]. With personalized medicine on the horizon, it is unquestionable that SNPs outside of the coding regions will be recognized for their importance in determining gene expression, which can then be used to predict phenotypic differences between individuals. An important application of the ChIP-seq method will be in the study of cancer cells, where the regulation of genes plays a fundamental part in the disease condition [50]. While the regulation of genes through methylation has been shown to play an important role in the development of cancer cells [49], the potential for cancerous cells to deregulate gene expression through the loss of existing transcription factor binding sites or the development of *de novo* sites through random sequence errors exists. It can be expected that ChIP-seq and its descendants will make these studies accessible, as the number and the quality of antibodies for transcription factors increase.

A consequence of cross-linking used in most ChIP-seq experiments is the capture of secondary interactions between the protein of interest and other proteins, which are, in turn, also cross-linked to other genomic locations. Because the chromosomes are able to adopt tertiary structures that bring seemingly distant regions into close contact, it is possible to obtain secondary protein–protein and protein–DNA interactions that may not be directly adjacent along the linear visualization of the genome. Mapping and identifying these long-range DNA interactions has the potential to contribute significantly to our understanding of the regions involved in gene expression and the mechanisms of transcriptional control. Two methods have been demonstrated using massively parallel sequencing and ChIP or ChIP-like approaches. Ruan and colleagues have recently demonstrated the ability to observe these interactions by adapting paired end tag strategies to identify long-range interactions in a method they unveiled in July 2007 called 454-Chia-PET. With this method, they were able to map the long-range genome-wide DNA interactions associated with estrogen receptor binding (Y. Ruan, personal communication). The second technology, called Chromosome Conformation Capture (3C) technology, applies formaldehyde cross-linking followed by a random digestion and a selective ligation step favoring cross-linked DNA fragments to capture sites of interacting DNA [60]. Both these processes have led to a series of ligated DNA fragment pairs representing interacting regions of the genome, which can be interpreted through the use of a paired end protocol, providing a genome-wide snapshot of the tertiary structure of chromosomes.

References

- 1 Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
- 2 Ptashne, M. and Gann, A. (1997) Transcriptional activation by recruitment. *Nature*, **386**, 569–577.
- 3 Reik, W., Romer, I., Barton, S.C., Surani, M.A., Howlett, S.K. and Klose, J. (1993) Adult phenotype in the mouse can be affected by epigenetic events in the early embryo. *Development*, **119**, 933–942.
- 4 Cheung, P. and Lau, P. (2005) Epigenetic regulation by histone methylation and histone variants. *Molecular Endocrinology*, **19**, 563–573.
- 5 ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
- 6 Strahl, B.D., Ohba, R., Cook, R.G. and Allis, C.D. (1999) Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 14967–14972.
- 7 Vignali, M., Hasan, A.H., Neely, K.E. and Workman, J.L. (2000) ATP-dependent chromatin-remodeling complexes. *Molecular and Cellular Biology*, **20**, 1899–1910.
- 8 Verdone, L., Agricola, E., Caserta, M. and Di Mauro, E. (2006) Histone acetylation in gene regulation. *Briefings in Functional Genomics and Proteomics*, **5**, 209–221.
- 9 Brutlag, D., Schlehuber, C. and Bonner, J. (1969) Properties of formaldehyde-treated nucleohistone. *Biochemistry*, **8**, 3214–3218.
- 10 Elnitski, L., Jin, V.X., Farnham, P.J. and Jones, S.J. (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Research*, **16**, 1455–1464.
- 11 Wolffe, A.P. (2001) Transcriptional regulation in the context of chromatin structure. *Essays in Biochemistry*, **37**, 45–57.
- 12 Orlando, V. (2000) Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends in Biochemical Sciences*, **25**, 99–104.
- 13 Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
- 14 Kuo, M.H. and Allis, C.D. (1999) *In vivo* cross-linking and immunoprecipitation for studying dynamic protein:DNA associations in a chromatin environment. *Methods*, **19**, 425–433.
- 15 Huang, H.S., Matevosian, A., Jiang, Y. and Akbarian, S. (2006) Chromatin immunoprecipitation in postmortem brain. *Journal of Neuroscience Methods*, **156**, 284–292.
- 16 Strahl-Bolsinger, S., Hecht, A., Luo, K. and Grunstein, M. (1997) SIR2 and SIR4 interactions differ in core and extended telomeric heterochromatin in yeast. *Genes & Development*, **11**, 83–93.
- 17 Hecht, A., Strahl-Bolsinger, S. and Grunstein, M. (1996) Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. *Nature*, **383**, 92–96.
- 18 Weinmann, A.S., Bartley, S.M., Zhang, T., Zhang, M.Q. and Farnham, P.J. (2001) Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Molecular and Cellular Biology*, **21**, 6820–6832.
- 19 LeBaron, M.J., Xie, J. and Rui, H. (2005) Evaluation of genome-wide chromatin library of Stat5 binding sites in human breast cancer. *Molecular Cancer*, **4**, 6.
- 20 Ahmadian, A., Ehn, M. and Hober, S. (2006) Pyrosequencing: history,

- biochemistry and future. *Clinica Chimica Acta*, **363**, 83–94.
- 21 Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
 - 22 Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002) Using the transcriptome to annotate the genome. *Nature Biotechnology*, **20**, 508–512.
 - 23 Roh, T.Y., Ngau, W.C., Cui, K., Landsman, D. and Zhao, K. (2004) High-resolution genome-wide mapping of histone modifications. *Nature Biotechnology*, **22**, 1013–1016.
 - 24 Chen, J. and Sadowski, I. (2005) Identification of the mismatch repair genes PMS2 and MLH1 as p53 target genes by using serial analysis of binding elements. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4813–4818.
 - 25 Ng, P., Wei, C.L., Sung, W.K., Chiu, K.P., Lipovich, L., Ang, C.C., Gupta, S., Shahab, A., Ridwan, A., Wong, C.H., Liu, E.T. and Ruan, Y. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods*, **2**, 105–111.
 - 26 Lin, C.Y., Vega, V.B., Thomsen, J.S., Zhang, T., Kong, S.L., Xie, M., Chiu, K.P., Lipovich, L. *et al.* (2007) Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genetics*, **3**, e87.
 - 27 Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
 - 28 Chiu, K.P., Wong, C.H., Chen, Q., Ariyaratne, P., Ooi, H.S., Wei, C.L., Sung, W.K. and Ruan, Y. (2006) PET-Tool: a software suite for comprehensive processing and managing of paired-end diTag (PET) sequence data. *BMC Bioinformatics*, **7**, 390.
 - 29 Drmanac, R., Drmanac, S., Chui, G., Diaz, R., Hou, A., Jin, H., Jin, P., Kwon, S. *et al.* (2002) Sequencing by hybridization (SBH): advantages, achievements and opportunities. *Advances in Biochemical Engineering/Biotechnology*, **77**, 75–101.
 - 30 Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, **98**, 503–517.
 - 31 Southern, E. (2006) Southern blotting. *Nature Protocols*, **1**, 518–525.
 - 32 Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R.A. and Dynlacht, B.D. (2002) E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes*, **16**, 245–256.
 - 33 Bernstein, B.E., Meissner, A. and Lander, E.S. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.
 - 34 Buck, M.J. and Lieb, J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.
 - 35 Draghici, S., Khatri, P., Eklund, A.C. and Szallasi, Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics*, **22**, 101–109.
 - 36 Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. and Kohane, I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.
 - 37 Jenssen, T.K., Langaas, M., Kuo, W.P., Smith-Sørensen, B., Myklebost, O. and Hovig, E. (2002) Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Research*, **30**, 3235–3244.
 - 38 Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M. and Brown, P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
 - 39 Fields, S. (2007) Site-seeing by sequencing *Science*, **316**, 1441–1442.
 - 40 Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide

- mapping of *in vivo* protein–DNA interactions. *Science*, **316**, 1497–1502.
- 41 Melamede, R.J. (1985–1989) Automatable process for sequencing nucleotide. US Patent 4,863,849.
 - 42 Leamon, J.H., Lee, W.L., Tartaro, K.R., Lanza, J.R., Sarkis, G.J., deWinter, A.D., Berka, J., Weiner, M. *et al.* (2003) A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, **24**, 3769–3777.
 - 43 Bennett, S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.
 - 44 Ng, P., Tan, J.J., Ooi, H.S., Lee, Y.L., Chiu, K.P., Fullwood, M.J., Srinivasan, K.G., Perbost, C. *et al.* (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Research*, **34**, e84.
 - 45 Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, **4**, 651–657.
 - 46 Reamon-Buettner, S.M. and Borlak, J. (2007) A new paradigm in toxicology and teratology: altering gene activity in the absence of DNA sequence variation. *Reproductive Toxicology*, **24**, 20–30.
 - 47 Morris, B.J. (2005) A forkhead in the road to longevity: the molecular basis of lifespan becomes clearer. *Journal of Hypertension*, **23**, 1285–1309.
 - 48 Feinberg, A.P. (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature*, **447**, 433–440.
 - 49 Herman, J.G. and Baylin, S.B. (2000) Promoter-region hypermethylation and gene silencing in human cancer. *Current Topics in Microbiology and Immunology*, **249**, 35–54.
 - 50 Jones, P.A. and Baylin, S.B. (2007) The epigenomics of cancer. *Cell*, **128**, 683–692.
 - 51 Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.
 - 52 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
 - 53 Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, **29**, 2994–3005.
 - 54 Gertz, E.M., Yu, Y.K., Agarwala, R., Schaffer, A.A. and Altschul, S.F. (2006) Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biology*, **4**, 41.
 - 55 Slater, G.S. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
 - 56 Lasko, T.A., Bhagwat, J.G., Zou, K.H. and Ohno-Machado, L. (2005) The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, **38**, 404–415.
 - 57 Reik, W., Dean, W. and Walter, J. (2001) Epigenetic reprogramming in mammalian development. *Science*, **293**, 1089–1093.
 - 58 Clark, S.J., Harrison, J., Paul, C.L. and Frommer, M. (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Research*, **22**, 2990–2997.
 - 59 Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
 - 60 Simonis, M., Kooren, J., and de Laat, W. (2007) An evaluation of 3C-based methods to capture DNA interactions. *Nature Methods*, **4**, 895–901.

16

MicroRNA Discovery and Expression Profiling using Next-Generation Sequencing

Eugene Berezikov and Edwin Cuppen

16.1

Background on miRNAs

MicroRNAs (miRNAs) are ~22 nt long RNA molecules, derived from genome-encoded stem-loop precursors, which can recognize target mRNAs by base pairing, and thereby regulate their expression. miRNAs were first discovered as regulators of developmental timing in *Caenorhabditis elegans* [1, 2]. However, it soon became obvious that miRNAs are an abundant class of small RNAs that are found in almost all organisms [3–5]. Now, the role of miRNAs in diverse developmental processes and disease is increasingly recognized, and specific miRNAs have been identified to play key roles in a variety of developmental and physiological processes (reviewed in Ref. [6]). miRNAs are transcribed from the genomic DNA as long transcripts, mostly by RNA polymerase II, and can exist in mono- or polycistronic configurations. The miRNA-coding genes are located in both intragenic (intronic) and intergenic regions [7]. The primary transcript, named pri-miRNA, folds into complex secondary structures (Figure 16.1), and is processed by the Drosha/Pasha nuclease complex in the nucleus. The resulting hairpin structure, called the pre-miRNA, has a 2 nt 3' overhang and is exported to the cytosol by the Exportin-5 complex, where it is further processed by the Dicer nuclease complex, which leaves 2 nt 3' overhangs, resulting in a duplex RNA structure consisting of the mature miRNA molecule and the star sequence (miRNA*). From the miRNA/miRNA* duplex, one strand, the miRNA, preferentially enters the protein complex that represses target gene expression – the Argonaute-containing RNA-induced silencing complex (RISC) – whereas the other strand is degraded. The choice of strand relies on the local thermodynamic stability of the miRNA/miRNA* duplex – the strand whose 5' end is less stably paired is loaded into the RISC [8, 9], although this process is not completely understood yet. The miRNA guides the RISC complex to target mRNA transcripts, where target sequences in the 3' UTR are supposedly recognized primarily by seed sequence (nucleotides 2–8 of the miRNA) pairing. Although a variety of computational tools and databases have been developed for the prediction of potential targets for

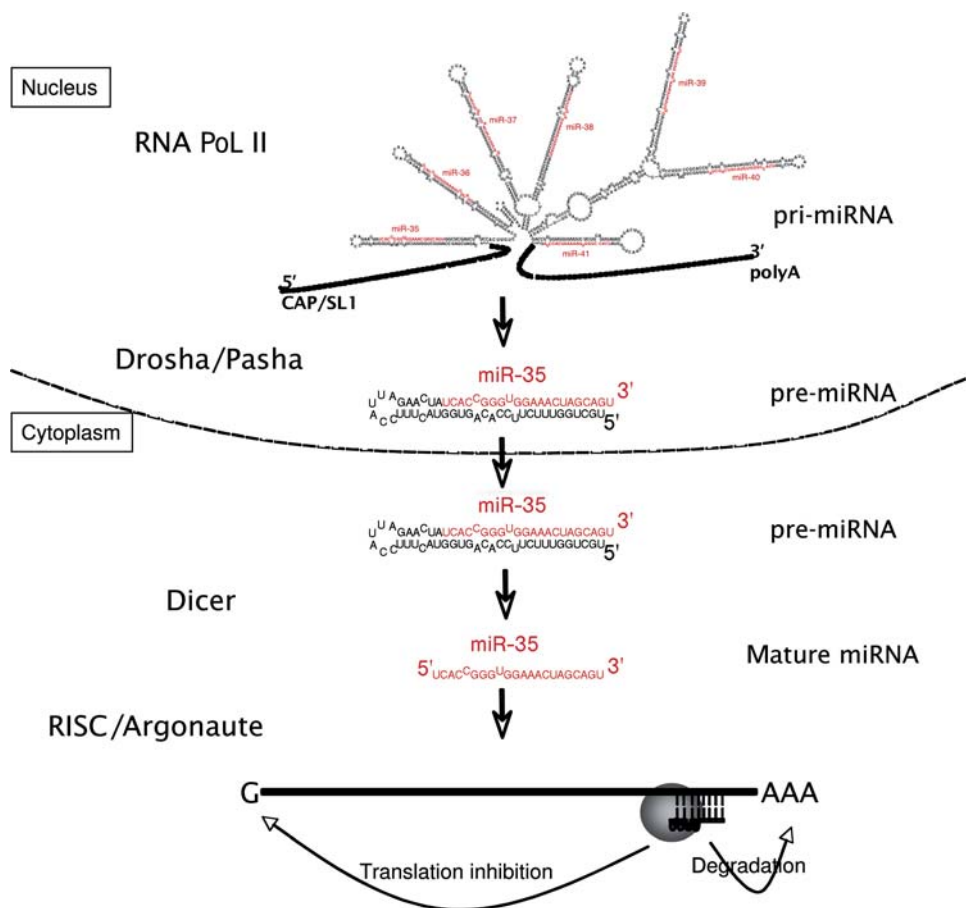


Figure 16.1 Model of the miRNA pathway. miRNA genes (e.g., *C. elegans* miR-35) are transcribed by RNA polymerase II. The transcripts fold into complex secondary structure (pri-miRNA). These structures are cropped by a complex containing Drosha and Pasha into hairpin precursors (pre-miRNA), which are

exported from the nucleus. In the cytoplasm, Dicer cleaves the pre-miRNAs and releases the mature miRNA, which is bound by the Argonaute proteins and guides them to the target mRNA. Upon targeting, translation is inhibited, or the mRNA is deadenylated and degraded.

miRNAs, experimental validation is required to demonstrate a genuine regulatory miRNA–mRNA relationship. Current estimates indicate that 30–50% of all genes in the human genome can be regulated by miRNAs.

16.2 miRNA Identification

Several hundred miRNA genes have initially been identified by sequencing of size-fractionated small RNA libraries in human and other vertebrates [3–5], and compu-

tational analyses have indicated that there could be substantially more miRNAs in the human genome [10, 11]. Although a variety of techniques were developed to validate computational predictions [10, 12], sequencing a large amount of cloned small RNAs has proven to be the most versatile and unbiased approach for miRNA discovery. It should be mentioned, however, that computational tools are still indispensable for the classification of miRNAs obtained in large-scale experimental efforts (see below).

Initially, cDNA libraries from small RNA fractions were made in a conventional way by cloning inserts in bacterial plasmid vectors, but with the availability of next-generation sequencing technology that allows single-molecule amplification and sequencing without cloning, exploration of the small RNA component of cells and tissues has become more simple and scalable (Figure 16.2; reviewed in Ref. [13]). As a result, applications of deep-sequencing approaches are no longer limited to expensive fishing expeditions but can now also be used for expression profiling of small RNA in routine experiments. The major advantages are that high-throughput sequencing allows the simultaneous (i) detection of known molecules and discovery of novel RNAs, thereby excluding the need for a priori knowledge on the molecules to be detected, and (ii) digital quantification of the number of molecules of a known or novel miRNA gene within a sample.

16.3

Experimental Approach

Experimentally, the following steps can be distinguished in a small RNA sequencing experiment using next-generation sequencing technology:

1. Sample collection and RNA isolation
2. Small RNA library construction
3. Sequencing
4. Bioinformatic analysis.

16.3.1

Sample Collection

The standard way to collect the small RNA fraction from tissue samples is to first isolate total RNA by using a protocol that is efficient in small RNA retainment (e.g., Trizol) or to immediately enrich the small RNA fraction (e.g., Ambion mirVana miRNA isolation). This step is followed by separation of the samples on a preparative polyacrylamide gel where the desired size range is excised (usually between 15 and 30 nt) followed by small RNA elution. The most common issues related to specific sample collection are essentially the same as for preparing samples for microarrays. Most important practical issues are the heterogeneity of the sample – for example, tissue or tumor samples consisting of multiple cell types and mixes of normal and cancer cells, respectively – and the difficulty to obtain reproducible tissue regions from different organisms for comparison. The good thing, however, is that miRNAs

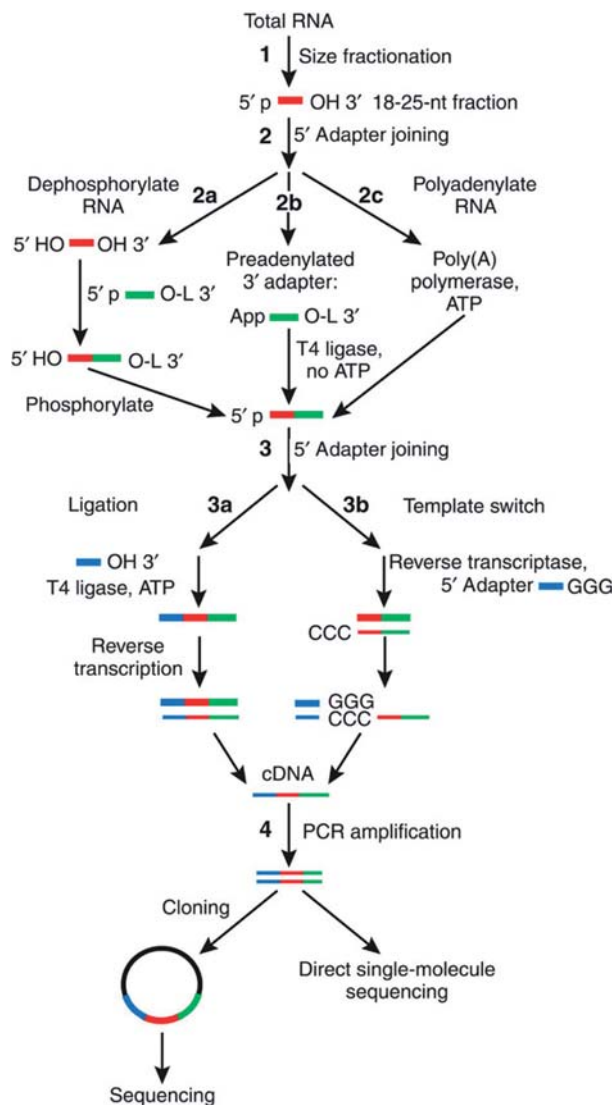


Figure 16.2 Strategies for cloning small RNAs. In 1, total RNA is separated on a polyacrylamide gel and the fraction corresponding to RNAs of 18–25 nt is recovered. In 2, a 3' adapter can be introduced in different ways: the adapter can be ligated to a dephosphorylated RNA, which is then phosphorylated (2a); a preadenylated adapter can be ligated to RNA without free ATP in the reaction (2b); or the RNA can be polyadenylated

by poly(A) polymerase (2c). In 3, a 5' adapter is introduced either by ligation (3a) or by template switching during reverse transcription (3b). In 4, cDNA is amplified by PCR and cloned into a vector to create a library. Alternatively, PCR products can be sequenced directly by single-molecule sequencing methods (massive parallel sequencing). Reprinted from [13].

are relatively stable in tissues and upon isolation, although one should still take care to isolate high-quality RNA as degraded large RNA could in principle be degraded to miRNA-sized fragments.

Since miRNAs are associated with Argonaute proteins, it is possible to perform enrichment of small RNAs for cloning by immunoprecipitation of Ago-containing complexes [14] or chromatographic isolation of miRNP complexes [15]. These methods have been developed recently and have the potential to substantially improve approaches for miRNA expression profiling.

16.3.2

Library Construction

Starting from small RNA, there are several protocols that can be followed to generate cDNA and to introduce the adapters at the 5' and 3' ends that are required for amplification and sequencing (Figure 16.2). For the first-strand cDNA synthesis, a 3' adapter needs to be ligated to the mature miRNA to introduce a site at which to anneal the primer used by reverse transcriptase. To prevent self-circularization of the mature miRNAs and the adapter, small RNAs are usually dephosphorylated before ligation and the 3'-hydroxyl terminus of the 3' adapter is blocked by incorporating a nonnucleotide group during chemical synthesis of the oligonucleotide [16]. In another popular variation of the protocol, the 3' adapter is preadenylated, removing the need to dephosphorylate the small RNA [4, 17]. Alternatively, ligation of the 3' adapter can be replaced by the addition of a poly(A) tail to the small RNAs using poly(A) polymerase [18, 19], in which case oligo(dT) is used as a primer for reverse transcription. In this case, there is also no need for dephosphorylation of the small RNA sample.

Before the reverse transcription reaction is performed, a 5' adapter is ligated to the gel-purified and, if necessary, rephosphorylated product of the 3' adapter ligation. Ligation of the 5' adapter can be omitted in protocols using cDNA cloning by SMART technology (Clontech) [5, 20], which relies on the property of specific types of reverse transcriptases to add several nontemplated nucleotides (predominantly deoxycytidine) to the 3' ends of synthesized cDNAs. These overhanging nucleotides can be subsequently used in switching the template from miRNA to the 5' adapter.

Traditionally, cDNA fragments were cloned in bacterial vectors, followed by plasmid-based sequencing of inserts [21], and the most extensive effort to date to characterize miRNA profiles in more than 250 different samples [22] followed this approach. Currently, massively parallel sequencing technology is rapidly replacing traditional plasmid sequencing, providing both increased sequencing depth and the omission of the plasmid-cloning step. The first example of using massively parallel sequencing for miRNA analysis was provided by Lu *et al.* [23] who applied massive parallel signature sequencing (MPSS) to elucidate the small RNA component of the *Arabidopsis thaliana* transcriptome. This technology enables hundreds of thousands of short (17 nt) sequencing tags to be generated in one run. Although this approach provides sufficient information for (small RNA) expression profiling, the discovery of novel small RNAs is limited by the cloned insert length of only 17 nt. Therefore, all

following studies used protocols that resulted in the cloning of the complete mature small RNA sequence.

Making a good small RNA “library” is still pretty laborious because of the different steps that require size selection and purification of samples from polyacrylamide gels. In addition, relatively a large amount of small RNA (>1 µg small RNA or >10 µg total RNA) is needed as input, although some protocols can use less input material [24].

Finally, a more difficult problem to solve is that some miRNAs may be hard to clone owing to their physical properties, including sequence composition and secondary structure, or posttranscriptional modifications, such as editing or methylation [25–27]. Cloning biases in small RNA library construction have indeed been reported in the literature [22]. It should be mentioned that several commercial kits for labeling small RNA for microarray applications do rely on adapter ligation or small RNA extension and may thus be equally sensitive to such modifications.

16.3.3

Massively Parallel Sequencing

At present, there are three different platforms commercially available (see other chapters in this book) that can be used for massively parallel sequencing of small RNA samples. The Roche/454 platform was the first commercially available system and most miRNA sequencing studies so far have been performed on it. However, the Illumina/Solexa and Applied Biosystems platforms have characteristics that better fit the specific demands for small RNA discovery and expression profiling – as many independent reads as possible are preferred and reads do not have to be longer than the typical length of a miRNA, which is less than 25 nt.

However, to maximally benefit from the relatively long-read lengths of the Roche/454 platform, small RNA cDNAs can be concatamerized into larger fragments before being flanked with amplification and sequencing adapters. This approach was originally successfully used to clone small RNAs into vectors to increase the length of informative sequence obtained from each sequenced clone [3, 4, 16]. Recently, a serial analysis of gene expression (SAGE)-like variation of the concatamerization step has been developed [28] that increases the average number of small RNA tags per clone from 5 to 35, thereby boosting the throughput and cost-efficiency of sequencing small RNA libraries.

Another way to most optimally benefit from the enormous amounts of sequencing reads that can be produced is the introduction of different barcodes in the adapters of different samples, followed by pooled sequence analysis. We have successfully used this approach to pool up to seven samples (unpublished results), but in principle there is no real limit to the depth of pooling. However, one should keep in mind that sequencing length may become limited to completely cover both the barcode and the small RNA sequence. This is especially an issue with the current read lengths of maximum 35 nt for the Illumina/Solexa and Applied Biosystems platforms. Alternatively, hybridization methods to decode beads or colonies that precede the sequencing reaction could theoretically be used to circumvent this problem but no

data on the successful application of this approach are available yet. When using barcodes in experiments with a quantitative nature, one will have to make sure that all barcoded adapters ligate equally well and do not introduce any cloning bias and therefore expression differences by themselves.

16.3.4

Bioinformatic Analysis

16.3.4.1 MicroRNA Discovery

Which cloned small RNA molecules are miRNAs? Answering this question is undoubtedly the most difficult part of the procedure. Although one could map back every individual read to the genome sequence and rely on annotation, such an approach would preclude the discovery of novel miRNAs and success depends very much on the quality and extent of genome annotation of the species that is studied. Other approaches are therefore needed, but these depend heavily on how a miRNA is defined. Previously, miRNAs were defined as noncoding RNAs that fulfill a combination of expression and biogenesis criteria [29]:

1. Mature miRNA should be expressed as a distinct ~22 nt transcript detectable by Northern blot analysis or other experimental means such as cloning from size-fractionated small RNA libraries.
2. Mature miRNA should originate from a precursor with a characteristic secondary structure – a hairpin – or fold-back, which does not contain large internal loops or bulges. Mature miRNA should occupy the stem part of the hairpin.
3. Mature miRNA sequences and predicted hairpin structure of the precursor should be conserved in different species.
4. Mature miRNA should be processed by Dicer as evidenced by increased precursor accumulation in Dicer-deficient mutants.

An “ideal” miRNA would meet all the above criteria. In practice, variations are possible; but at the very minimum, to classify a novel sequence as a miRNA, the expression of a ~22 nt form and the presence of a hairpin precursor need to be demonstrated. Although it may seem a trivial task to determine the genomic location (or locations) of a 22-nt sequence and to check whether a phylogenetically conserved hairpin precursor is encoded in the genomic region, such analysis is complicated, however, by the fact that hairpin structures are common in eukaryotic genomes and are not a feature unique to miRNAs. Moreover, mature miRNA molecules often have nontemplated nucleotides at their 3' end [22], complicating mapping of the reads to genome.

With the advent of next-generation sequencing, several groups, including ours, have developed sophisticated computational pipelines to discover novel miRNAs in deep sequencing data sets [19, 30–33]. These approaches follow largely the same principles and aim to incorporate most of the characteristics needed to qualify a small RNA as a miRNA (Figure 16.3). First, the genomic locus from which the molecule is derived should be able to produce an RNA precursor that folds in a stable hairpin structure that is an optimal substrate for Drosha and Dicer nucleases,

as judged from hairpin pairing characteristics. Second, the small RNA should have solid evidence for expression, for example, supported by multiple clones and/or observed in several independent libraries. Third, cloning of the “by-product” of Drosha/Dicer processing, the star sequence, is considered a strong evidence supporting miRNA biogenesis of the molecule. At the same time, the presence of the star sequence that does not show the characteristic 2-nt 3′ overhang in the duplex can be used to “disqualify” the molecule as a miRNA. Finally, the annotation of the genomic region in question should not suggest non-miRNA origin of the molecule (e.g., overlap with tRNA or rRNA genes, protein-coding sequences or repeats). In addition to these four criteria, phylogenetic conservation of the sequence with typical camel-like conservation profile [11, 34] over the pre-miRNA region adds strong support for a genuine miRNA, although this argument does not necessarily have to be true as species-specific miRNAs have been described [10, 35]. Furthermore, additional support could be found within a species when novel candidate miRNAs are related (seed sequence identity) to known miRNAs and form a subfamily. Likewise, miRNAs are often located in genomic clusters [7] and expressed from long precursor RNAs.

16.3.4.2 miRNA Expression Profiling

Cloning frequencies of miRNAs in principle should reflect their abundance in the source tissue, and indeed several studies have shown that deep sequencing data faithfully recapitulate miRNA expression patterns [22, 31, 32, 36]. The main question in the interpretation of miRNA expression profiles derived from clone counts is how data can be best normalized before comparing different samples? It is possible to use for normalization either (a) the total number of small RNA clones or (b) the number of sequenced ribosomal/tRNA clones or (c) a subset of specific miRNA clones or (d) the total number of miRNA clones. While the total number of small RNA clones is used in some studies [15], Ruby *et al.* [32] report that normalization to the total number of miRNA clones produces best correlation to published miRNA profiles generated by Northern blot analysis. Normalization to the total number of miRNAs can be rationalized by the assumption that different cells contain approximately the same amount of RISC complexes and that most miRNA molecules in the cell are present in the RISC-bound state.

16.4

Validation

The question remains when candidate miRNA can be qualified as a genuine miRNA. Independent validation using, for example, Q-PCR or Northern blot can help to further delineate expression (levels) of the candidate, but such experiments do not prove whether the detected molecule is a product of Dicer processing or whether the molecule regulates target mRNAs. Unfortunately, current functional assays are extremely laborious, making them difficult to parallelize and scale up. Hence, novel approaches need to be developed to effectively address this issue.

16.5

Outlook

The approaches described in this chapter are not limited to miRNAs, but they also identify other classes of small RNAs. The only limitations are that the size and cloning strategy are compatible with the characteristics (e.g., length, 5' and/or 3' modification status) of the target molecule. Recently, piRNA and mirtron genes were identified [33, 37–41], and experimental support largely stems from molecules sequenced by massively parallel sequencing approaches. More classes of small RNAs may be discovered by using such approaches, although it is clear that miRNAs are by far the most abundant class of small RNA molecules.

Finally, the question remains whether sequence-based expression profiling will replace microarray-based approaches not only for miRNA detection but also for mRNA profiling [42]. Although the quality of the resulting data should be the driving force behind this change, cost aspects will undoubtedly play a role as well. However, costs per read for massively parallel sequence have already been going down gradually, in part because of increased numbers of reads per run and decreasing running costs, which may further decrease with increased competition in the next- and future-generation sequencing market.

References

- 1 Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, **75**, 843–854.
- 2 Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R. and Ruvkun, G. (2000) The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, **403**, 901–906.
- 3 Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
- 4 Lau, N.C., Lim, L.P., Weinstein, E.G. and Bartel, D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.
- 5 Lee, R.C. and Ambros, V. (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.
- 6 Kloosterman, W.P. and Plasterk, R.H.A. (2006) The diverse functions of microRNAs in animal development and disease. *Developmental Cell*, **11**, 441–450.
- 7 Kim, V.N. and Nam, J.W. (2006) Genomics of microRNA. *Trends in Genetics*, **22**, 165–173.
- 8 Khvorova, A., Reynolds, A. and Jayasena, S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.
- 9 Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N. and Zamore, P.D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, **115**, 199–208.
- 10 Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y. and Bentwich, Z. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics*, **37**, 766–770.
- 11 Berezikov, E., Guryev, V., van de, B.e.J., Wienholds, E., Plasterk, R.H. and

- Cuppen, E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.
- 12 Berezikov, E., van Tetering, G., Verheul, M., van de Belt, J., van Laake, L., Vos, J., Verloop, R., van de Wetering, M., Gurjev, V., Takada, S., van Zonneveld, A.J., Mano, H., Plasterk, R. and Cuppen, E. (2006) Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Research*, **16**, 1289–1298.
 - 13 Berezikov, E., Cuppen, E. and Plasterk, R.H.A. (2006) Approaches to microRNA discovery. *Nature Genetics*, **38** (Suppl), S2–S7.
 - 14 Nelson, P.T., De Planell-Saguer, M., Lamprinak, S., Kiriakidou, M., Zhang, P., O'Doherty, U. and Mourelatos, Z. (2007) A novel monoclonal antibody against human Argonaute proteins reveals unexpected characteristics of miRNAs in human blood cells. *RNA*, **13**, 1787–1792.
 - 15 Gu, S.G., Pak, J., Barberan-Soler, S., Ali, M., Fire, A. and Zahler, A.M. (2007) Distinct ribonucleoprotein reservoirs for microRNA and siRNA populations in *C. elegans*. *RNA*, **13**, 1492–1504.
 - 16 Pfeffer, S., Lagos-Quintana, M. and Tuschl, T. (2005) *Current Protocols in Molecular Biology* (eds F. Ausubel, R. Brent, R. Kingston, D. Moore, J. Seidman, J. Smith and K. Struhl), Wiley, San Francisco pp. 26.4.1–26.4.18.
 - 17 Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F.A., van Dyk, L.F., Ho, C.K., Shuman, S., Chien, M., Russo, J.J., Ju, J., Randall, G., Lindenbach, B.D., Rice, C.M., Simon, V., Ho, D.D., Zavolan, M. and Tuschl, T. (2005) Identification of microRNAs of the herpesvirus family. *Nature Methods*, **2**, 269–276.
 - 18 Fu, H., Tie, Y.i., Xu, C., Zhang, Z., Zhu, J., Shi, Y., Jiang, H., Sun, Z. and Zheng, X. (2005) Identification of human fetal liver miRNAs by a novel method. *FEBS Letters*, **579**, 3849–3854.
 - 19 Berezikov, E., Thuemmler, F., van Laake, L., Kondova, I., Bontrop, R., Cuppen, E. and Plasterk, R.H. (2006) Diversity of microRNAs in human and chimpanzee brain. *Nature Genetics*, **38**, 1375–1377.
 - 20 Takada, S., Berezikov, E., Yamashita, Y., Lagos-Quintana, M., Kloosterman, W.P., Enomoto, M., Hatanaka, H., Fujiwara, S.-i., Watanabe, H., Soda, M., Choi, Y.L., Plasterk, R.H.A. and Cuppen, E. (2006) Mouse microRNA profiles determined with a new and sensitive cloning method. *Nucleic Acids Research*, **34**, e115.
 - 21 Meyers, B.C., Souret, F.F., Lu, C. and Green, P.J. (2006) Sweating the small stuff: microRNA discovery in plants. *Current Opinion in Biotechnology*, **17**, 139–146.
 - 22 Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., Lin, C., Socci, N.D., Hermida, L., Fulci, V., Chiaretti, S., Foa, R., Schliwka, J., Fuchs, U., Novosel, A., Muller, R. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
 - 23 Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C. and Green, P.J. (2005) Elucidation of the small RNA component of the transcriptome. *Science*, **309**, 1567–1569.
 - 24 Mano, H. and Takada, S. (2007) mRAP, a sensitive method for determination of microRNA expression profiles. *Methods*, **43**, 118–122.
 - 25 Luciano, D.J., Mirsky, H., Vendetti, N.J. and Maas, S. (2004) RNA editing of a miRNA precursor. *RNA*, **10**, 1174–1177.
 - 26 Yang, W., Chendrimada, T.P., Wang, Q., Higuchi, M., Seeburg, P.H., Shiekhattar, R. and Nishikura, K. (2006) Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nature Structural & Molecular Biology*, **13**, 13–21.
 - 27 Yang, Z., Ebright, Y.W., Yu, B. and Chen, X. (2006) HEN1 recognizes 21–24 nt small RNA duplexes and deposits a methyl group onto the 2' OH of the 3' terminal nucleotide. *Nucleic Acids Research*, **34**, 667–675.

- 28 Cummins, J.M., He, Y., Leary, R.J., Pagliarini, R., Diaz, L.A., Jr., Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A.E., Labourier, E., Raymond, C.K., Roberts, B.S., Juhl, H., Kinzler, K.W., Vogelstein, B. and Velculescu, V.E. (2006) The colorectal microRNAome. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3687–3692.
- 29 Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G. and Tuschl, T. (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.
- 30 Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H. and Bartel, D.P. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, **127**, 1193–1207.
- 31 Calabrese, J.M., Seila, A.C., Yeo, G.W. and Sharp, P.A. (2007) RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 18097–18102.
- 32 Ruby, J.G., Stark, A., Johnston, W.K., Kellis, M., Bartel, D.P. and Lai, E.C. (2007) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Research*, **17**, 1850–1864.
- 33 Berezikov, E., Chung, W.J., Willis, J., Cuppen, E. and Lai, E.C. (2007) Mammalian mirtron genes. *Molecular Cell*, **28**, 328–336.
- 34 Berezikov, E. and Plasterk, R.H. (2005) Camels and zebrafish, viruses and cancer: a microRNA update. *Human Molecular Genetics*, **14**, R183–R190.
- 35 Berezikov, E., Thuemmler, F., van Laake, L.W., Kondova, I., Bontrop, R., Cuppen, E. and Plasterk, R.H. (2006) Diversity of microRNAs in human and chimpanzee brain. *Nature Genetics*, **38**, 1375–1377.
- 36 Lui, W.O., Pourmand, N., Patterson, B.K. and Fire, A. (2007) Patterns of known and novel small RNAs in human cervical cancer. *Cancer Research*, **67**, 6031–6043.
- 37 Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P. and Kingston, R.E. (2006) Characterization of the piRNA complex from rat testes. *Science*, **313**, 363–367.
- 38 Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R. and Hannon, G.J. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, **128**, 1089–1103.
- 39 Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den, E.I.H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B., Plasterk, R.H., Hannon, G.J., Draper, B.W. and Ketting, R.F. (2007) A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. *Cell*, **129**, 69–82.
- 40 Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M. and Lai, E.C. (2007) The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell*, **130**, 89–100.
- 41 Ruby, J.G., Jan, C.H. and Bartel, D.P. (2007) Intronic microRNA precursors that bypass Drosha processing. *Nature*, **448**, 83–86.
- 42 Torres, T.T., Metta, M., Ottenwalder, B. and Schlotterer, C. (2007) Gene expression profiling by massively parallel sequencing. *Genome Research*, **18**, 172–177.

17

DeepSAGE: Tag-Based Transcriptome Analysis Beyond Microarrays

Kåre L. Nielsen, Annabeth H. Petersen, and Jeppe Emmersen

17.1

Introduction

All living things carry their genetic information in genes usually in the form of DNA. The last decade has seen the completion of a multitude of genome sequences, most notably of *Escherichia coli* [1], yeast [2], *Arabidopsis thaliana* [3], rice [4, 5], mouse [6], and not the least, the human genome [7, 8]. Each constitutes a paramount achievement in molecular biology, and all are great contributions to our understanding of living organisms. However, the activity of genomes is regulated to meet the requirement by the organism itself, or as a response to external abiotic factors such as light, heat, and temperature, and to biotic factors such as infection by pathogens. Genes are transcribed into mRNAs that are, in turn, translated into proteins and catalytically active enzymes. Regulation of this system is primarily obtained by controlling the amount of mRNA that is produced from each gene and the turnover of the corresponding protein. The mRNA population is often referred to as the transcriptome. The complexity of the system is enormous; today, it is believed that about 25 000 genes are present in any genome of animals and higher plants.

Consequently, to understand the genetics that underlies biological change such as development, disease, crop yield, or resistance, knowing the genetic parts list is not sufficient; it has proven both informative and necessary to perform comparative transcriptomic studies to understand how genomes are regulated in response to biological change [9–11].

Gene expression analyses of single genes were performed by Northern blots [12] when the first high-throughput method for quantitatively determining gene expression of multiple genes, expressed sequence tags (ESTs), was developed in 1993 [13]. Today, the sequencing of ESTs still provides great value for money in terms of obtaining information on gene sequences, but a severe bias against small and large transcripts during EST analysis limits its reliability and hence its usefulness for quantitative transcriptome analysis [14–16].

In the mid-1990s, almost simultaneously two entirely new transcriptome analysis methods were presented: DNA microarrays [17] and serial analysis of gene expression (SAGE) [18]. They rely on the same two fundamentally different principles as Northern blots and EST sequencing. DNA microarray analysis, as Northern blots, uses complementary DNA hybridization and analogue quantification of intensity signals, whereas SAGE, as EST sequencing, relies on DNA sequencing and digital counting of specific sequence tags. However, two very important expansions were introduced; by massive paralleling the hybridization experiments on tiny spots on a DNA chip, hybridization became a high-throughput quantitative technique and by eliminating the size bias of ESTs by isolating small tags of the same length from transcripts, instead of cloning, and sequencing the entire transcript, SAGE became a representative as well as a high-throughput technique [15].

Although slightly a younger method of the two, the dominant method for high-throughput gene expression profiling today is DNA microarrays. Current arrays may consist of more than 100 000 unique single-stranded DNA molecules attached to a glass slide in an ordered fashion. Two samples of mRNA are prepared and labeled with two different fluorescence labels, mixed and hybridized to the array. At positions, where the amount of mRNA is different between the two samples, one of the two fluorescence signals is in excess. This is quantified and because the DNA sequence at a particular position is known to be unique to one mRNA, it provides a measure of the relative amount of mRNA present between the two samples. An advantage of DNA microarrays is that once the array has been made at a very high cost, many measurements can be made at a relatively low cost. However, due to the “analogue” nature of the signal, there is a limited dynamic range of measurements and quantitative hybridization experiments are very difficult to carry out reproducibly in practice. Furthermore, only known genes can be spotted on the array, so it requires a detailed knowledge of the genetic background. Recently, genome-tiling arrays encompassing the overlapping nucleotides for the entire genome sequence have enabled DNA microarrays to expand transcript analysis beyond annotated genes, but only for a few select organisms for which a whole genome array is available [19].

SAGE, on the other hand, can measure the expression of both known and unknown genes. It relies on the extraction of a unique 20 bp sequence (tag) from each mRNA. These tags are traditionally ligated together end-to-end and sequenced. By high-throughput DNA Sanger sequencing equipment, a typical sequence run of 96 samples about 1600 tags, and therefore mRNAs, can be detected. Typically, determining 50 000 tags can provide detailed knowledge of the 2000 most highly expressed genes in the tissue analyzed, and this is not limited to previously known genes. Therefore, SAGE can be used to discover new genes. Unknown tags obtained through SAGE analysis of a sample can be efficiently used as gene-specific primers in rapid amplification of cDNA ends (RACE) reactions to generate full-length transcripts that can be cloned and sequenced [20].

Furthermore, because SAGE is a “digital” method, the sensitivity is limited only by the amount of tags sampled and can therefore be extended beyond the sensitivity of

microarrays. This is an important feature in the exploration of transcriptomes, because it facilitates the reliable quantification of *the master gene regulators* – the transcription factors. While SAGE for these reasons constitutes an attractive alternative to microarrays, it has two major drawbacks: it is slow compared to microarrays and it is expensive due to the high cost of sequencing, and the manual labor or robotics needed for colony picking, library construction, and sample preparation.

However, taking advantage of the recent emergence of the new-generation sequencing technologies and their capacity to generate a very large number of short DNA reads [21, 22], DeepSAGE has been developed [23]. Using DeepSAGE decreases cost, complexity, and labor dramatically, and today it is now possible to harvest all the advantages of tag-based transcriptomics at a lower cost than microarray analysis.

17.2

DeepSAGE

A DeepSAGE analysis experiment consists of four parts: (i) collection of biological samples and extraction of high-quality total RNA; (ii) processing of mRNA into DNA tags flanked by appropriate linker sequences; (iii) sequencing of tags; and (iv) data analysis.

High quality and proper handling of the biological sample is of fundamental importance to any experiment. However, it is especially important for transcriptome studies. While RNA in chemical terms is a remarkably stable biological molecule, and in biological terms a rather abundant molecule, the abundance, processivity and robustness of RNA-degrading RNases frequently hinders the purification of high-quality RNA. The quality of RNA can be assessed either by using microfluidic devices such as Agilent's Bioanalyzer or by comparing ribosomal RNA band intensities in agarose gels. The principal determinant of success is efficient and quick lysis of the tissue in question resulting in quick inactivation of RNases present *in situ*. Therefore, to maximize the surface the lysis buffer can attack, grinding of fibrous and hard material (such as bone or cartilage) and cell wall containing tissues such as those from plants and fungi in liquid nitrogen is often necessary. This is costly and labor intensive, and obstructs high-throughput processing of such samples. Importantly, scaling down sample preparation of tissues not only minimizes cost but also increases the surface-to-volume ratio of tissue samples, thus decreasing the need for prelysis processing of the samples. In addition, some biological samples, such as medical biopsies, can be obtained only in very small amounts. For these reasons, the amount of RNA needed for transcriptome analysis is an important parameter that can potentially decrease cost and labor of transcriptome analysis studies and facilitate analysis of previously unattainable samples.

Dramatic decreases in amount of RNA needed for analysis have been obtained with modification to the traditional SAGE protocol. Velculescu used 5 µg of mRNA corresponding to approximately 1 mg of total RNA for the first SAGE study [18]. The current version of the protocol associated with the commercial kit available from

Invitrogen recommends the use of 5–50 µg of total RNA, a 200-fold decrease. However, it still corresponds to approximately 10^7 cells, which is a very high number for studies of rare cell types, such as stem cells or cancer cells obtained by FACS sorting. Employing preamplification of RNA, SAGE data from as little as 40 ng of total RNA obtained from 3000 microdissected cells have been reported [24]. Since PCR amplification of linker-flanked tags or ditags precedes DNA sequencing and generally can be performed without bias at this stage [25], the amount of material needed primarily depends on the amount of loss at each step in the protocol (see Figure 17.1). Because DeepSAGE contains fewer steps than SAGE, there is potential for reducing the amount of starting material even further (see Figure 17.1). In the current protocol for DeepSAGE in our laboratory, we use 2.5 µg of total RNA, but have data that suggest that we can go down to at least 200 ng without the need for preamplification (AHP and KLN, unpublished). In comparison, a standard Affymetrix microarray experiment uses 5 µg of total RNA as input material (www.affymetrix.com).

Figure 17.1 shows an overview of the LongSAGE procedure and two versions of the DeepSAGE procedure for the two different platforms, 454 and Solexa sequencing. In all procedures, mRNA is hybridized to poly-T-coated paramagnetic beads and double-stranded cDNA is synthesized. After the beads are washed, a frequent cutting restriction enzyme, typically *Nla*III, is used to cleave the cDNA. This often occurs at multiple sites depending on the sequence of individual transcripts, but only the bead-attached 3'-most fragment is retained. The *Nla*III overhang present on all bead-attached cDNA fragments is used to ligate linkers containing a recognition site for the type II enzyme, *Mme*I, which cleaves double-stranded DNA 20 nucleotide downstream to its recognition sequence, thus liberating a linker-tag molecule from the bead. In LongSAGE and DeepSAGE for 454 sequencing, these are ligated together to form linker-ditags. In DeepSAGE for Solexa sequencing, owing to the short-read lengths (35 nt) of Solexa, Linker-B is ligated directly onto the linker-tag to form a dilinker-monotag. These molecules can be amplified by PCR without introducing bias, because all linker-(di)tags are of the same length [25]. In LongSAGE, the linkers are removed by digestion with *Nla*III and purification of the ditag molecule, which is concatenated by ligation. The concatemers are sorted by size by gel electrophoresis, purified, joined to a plasmid vector, and a library of concatemer-containing recombinant bacteria is generated. Unfortunately, it is this latter part of the protocol that is the most difficult to succeed, and the process is terribly wasteful. Up to 400 PCR reactions of 50 µl are required to generate enough clones to facilitate the sequencing of about 50 000 tags that seems to have been adopted by the SAGE community as the cost-effective size of a SAGE study. Two phenomena reduce the yield of clones. While small DNA fragments are readily and quantitatively isolated from polyacrylamide gels, this is not the case for the much larger concatemers limiting the amount of purified, size-selected concatemers that can be used for library generation. Using agarose gels instead seems to decrease the efficiency of cloning for unknown reasons. In addition, the formation of DNA minicircles, which cannot be cloned, is a likely event and favored at lower free ditag concentration, which in turn, varies over time. It is therefore very difficult to optimize this very critical ligation reaction.

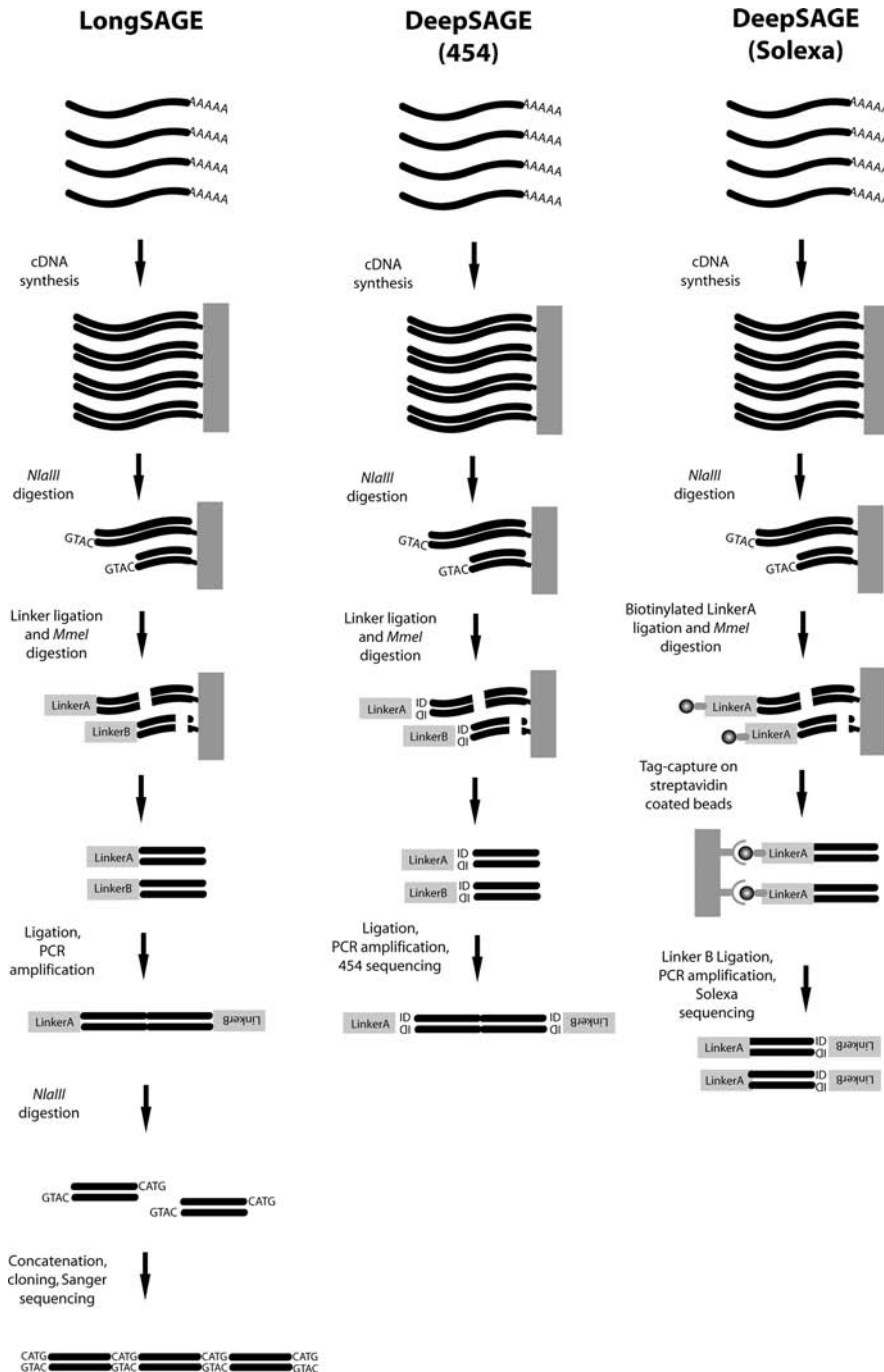


Figure 17.1 Overview of LongSAGE and DeepSAGE procedures for 454 and Solexa sequencing, respectively. ID keys in DeepSAGE that facilitate multiplexing of samples. Note that linker sequences are varying, since they are specific to sequencing platform.

Improvements have been suggested, for example [26], to limit the amount of material needed. But still, failure to obtain enough clones to complete the transcriptome study often requires starting all over again with the protocol. In addition, plating of the library, picking of clones, growing these individually, and preparing them for sequencing are not trivial procedures and great care should be taken to avoid cross-contamination of clones. In general, it requires the use of clone-picking robots (usually available only at sequencing centers) or quite a large manual effort. DeepSAGE circumvents these tedious and inefficient procedures by exploiting the ultrahigh-throughput and cloning-free sample preparation for 454 and Solexa sequencing. By sequencing the linker-ditags (DeepSAGE-454) or linker-monotags (DeepSAGE-Solexa) directly, a single PCR reaction of 25 μ l contains sufficient molecules to facilitate the analysis of millions of tags [23].

The 454 GS FLX (Roche) can provide up to 500 000 ditag sequences. From about 75% of these, a high-quality ditag sequence can be extracted. Therefore, 750 000 tags can be obtained in a single run. In comparison, the Solexa system routinely provides 25–30 million sequences, of which 75% contain a high-quality monotag sequence; consequently, 19–22 million tags are obtained from a single Solexa run (AHP, JE, and KLN, unpublished). Considering that the community standard for LongSAGE is 50 000 tags and it has been argued that 120–150 000 tags correspond to the sensitivity of DNA microarrays [27], DeepSAGE represents a much deeper sampling of the transcriptome than has previously been possible.

However, 454 and Solexa combined with DeepSAGE were not the first to facilitate the very deep analysis of the transcriptome. The now discontinued service, massive parallel signature sequencing (MPSS), was also capable of producing millions of tags in a single run, but the costs involved were very high [28]. This has resulted in relatively few studies published, although the techniques have been available since 2000 [29–33]. Recently, comparing data between MPSS and SAGE has shown that MPSS detects fewer transcripts despite the deeper sampling [34]. The reason for this is unknown, but it raises serious doubt about the reliability of results obtained by MPSS.

Owing to the high sequencing power of 454 and Solexa, it may well be cost effective to divide a sequence run over multiple samples. To this end, the DeepSAGE procedure includes the addition of variants of linker molecules that can be added to different biological samples [23]. Each of these contains a specific nucleotide identification key. This key is used in the downstream bioinformatical processing to sort the obtained sequences according to the biological samples from which they were derived without the need of physically keeping the samples apart while performing the sequencing. In theory, there is an unlimited amount of ID keys possible, since 256 different sequences are possible from a four-nucleotide key, 1024 sequences are available from a five-nucleotide key, and so on. In practice, however, the number of possible combinations is considerably smaller. It is desirable to include only those sequences that are more than two substitutions away from each other, so that sequencing errors cannot transform one ID key into another. Ultimately, the custom synthesis of a very large number of different adapters is cost prohibitive.

17.3

Data Analysis

Primary analysis of DeepSAGE data, such as extraction and counting of tags from sequence files, assessment of SAGE library quality, and filtering for low-quality sequences, can be performed by using the Perl scripts freely available from http://www.bio.aau.dk/en/biotechnology/software_applications/perl_scripts.htm (Figure 17.2).

The resulting data structure of SAGE is simple: a list of tag sequences each associated with a count, reflecting the number of times they occurred. However, the tags themselves are not very informative; only when tags are matched to biological sequence databases and associated with genes do they acquire interpretable biological information in terms of mapping directly not only to, for example, RefSeq transcript entries but also to biological functional databases like Gene Ontology (GO, www.geneontology.org) and Kyoto Encyclopedia of Genes and Genomes (KEGG, <http://www.genome.ad.jp/kegg/pathway.html>) (scripts for the mapping of tags can also be found at http://www.bio.aau.dk/en/biotechnology/software_applications/perl_scripts.htm).

A crucial point to consider, which directly influences data quality, is the reliability of which obtained tags can be uniquely matched to single genes. The original SAGE procedure used the restriction enzyme *BsmFI* to produce 11 nt tags. This was abandoned, because it was realized that 11 nt was insufficient to uniquely identify transcripts [35, 36]. LongSAGE, like DeepSAGE, uses *MmeI* to generate 17 nt tags and the resulting tags match a single gene in more than 90% of human genes [35]. In cases where a tag matches more than one gene, the genes are usually highly homologous (KLN, unpublished observation).

In contrast to microarrays, where intensity in a given spot depends on hybridization efficiency of the specific oligonucleotide sequence and the complementary transcript length and structure, transcript abundance detected by SAGE from one gene can readily be compared with transcript abundance from another gene. This is possible because all mRNA copies have the same chance of ending up in a linker-tag molecule. Only exception to this rule is the small percentage of transcripts that does not include a *NlaIII* recognition sequence; such transcripts are not detected at all. This limitation can be overcome by combining two frequent cutting restriction enzymes (e.g., *NlaIII*, *Sau3A*, and *DpnII*), which is feasible given the increase in sequencing throughput now available [35].

17.4

Comparing Tag-Based Transcriptome Profiles

The general reliability of SAGE to detect changes in the transcriptome has been compared with other profiling methods and generally found to be high [24, 37–40]. Typically 70–85% of gene expression profiles obtained by SAGE can be confirmed by different methods, such as DNA microarrays or RT-PCR [24, 40]. This proportion of

Tag sequence	Coun	Accession no.	Gene name
CATGAGCATCTCTGCATCT	2269	CN465657	Homologue to SPI P56168 MT21_B Metallothionein-like protein type 2 MT2-4/MT2-2
CATGAAGTGAAGTGAATGA	1023	CV502454	
CATGGATTTCTAAATAAAT	878	TC119014	Homologue to UPI SPI5_SOLTU (Q41484) Serine protease inhibitor 5 precursor (gCI
CATGTAGTTTCAATATCTG	780	CN462155	
CATGTAAATTAATGTATGC	532	Unknown	
CATGTTATGTTGTTAAAA	496	Unknown	
CATGACAACTGAACCTG	490	TC111928	Homologue to UPI TUB8_SOLTU (P33191) Induced stolen tip protein TUB8 (Fragme
CATGTAAATTAATGTATGC	483	TC111750	similar to UPI CPI8_SOLTU (O24384) Cysteine protease inhibitor 8 precursor (PCPI-
CATGTAATTTAAATGCTTA	476	Unknown	
CATGTTAGGTTATAATCT	453	Unknown	
CATGCATGGAGGTGCACAG	402	Unknown	
CATGGATTTCTAAAAAAA	390	Unknown	
CATGACTGATTATTACCTT	379	CN516452	Homologue to PIR T07592 T07 class I patatin - potato (Solanum tuberosum.);, partial
CATGATGTAATGTTTGT	367	CV504183	
CATGCCGCCACAAGAAACA	347	BF153386	
CATGGTTTGATTAAAGATG	334	TC111928	Homologue to UPI TUB8_SOLTU (P33191) Induced stolen tip protein TUB8 (Fragme
CATGTAATTTAATGTATGC	322	TC119013	UPI CPI9_SOLTU (Q00652) Cysteine protease inhibitor 9 precursor (PKIX) (pT1), p2
CATGGGATAGTTGCTGTGA	300	TC119141	UPI Q42897 (Q42897) Ubiquitin conjugating enzyme E2, complete
CATGTCGAGTATTATGTAG	288	Unknown	
CATGGTCGCTCCAGACCA	281	TC111996	Homologue to UPI Q6SKP4 (Q6SKP4) Ribosomal protein L3, complete
CATGAGTGCAAAAATATCG	275	Unknown	
CATGGTGGCTCAGGATTTT	271	Unknown	
CATGTTTACGCAATTTGTT	270	Unknown	
CATGCTAAAATATAGATAT	266	CN516429	
CATGAGAAGTGCTGAGAGT	260	CN465312	Homologue to UPI AAR83888 (AAR83888) Auxin-repressed protein ARP1, partial (7
CATGAGCTTTTGTCTCTTT	250	TC119057	UPI Q9M3H3 (Q9M3H3) Annexin p34, complete
CATGATGATATAACTGAGT	242	Unknown	
CATGCAAAATTTTGTCAAG	239	TC111734	TIGR_Osa1 9634.m04552 polyubiquitin - maize, complete
CATGTTTTTGGTTTCGTCA	237	TC111714	homologue to TIGR_Osa1 9639.m04467 dnaK-type molecular chaperone hsp70 - ric
CATGAAGGAAGCAATCTCA	232	CK851160	PIR T10479 T10 glycine-rich RNA-binding protein GRP1 - Commerson's wild
CATGGATTTCTAAATAAAA	221	Unknown	
CATGTGATGAAGGTATATA	219	TC119042	UPI PHS1_SOLTU (P04045) Alpha-1,4 glucan phosphorylase, L-1 isozyme, chloropl
CATGGCTTTCTAAAAAAA	215	Unknown	
CATGGCTTTCTAAAAAAT	215	TC111943	Homologue to UPI APIA_SOLTU (Q03197) Aspartic protease inhibitor 10 precursor (
CATGGACATTTGTGACGA	214	TC120880	similar to UPI Q42393 (Q42393) Sn-1 protein, partial (94%)
CATGTGTTTTGACATTCC	200	TC125991	UPI Q9ZRB6 (Q9ZRB6) C/21A protein, complete

tags that can be cross-validated is similar to those found for DNA microarrays [40]. The fact that the methods do not overlap completely is presumably because so far, given the sensitivity of existing methods, it has only been possible to sample a very small part of the transcriptome in either method. The differences observed are then likely to be small imperfections in the methods causing some transcripts to have slightly greater change of detection in one system compared to the other.

Assuming the overall reasonable fundamental statement that all mRNA copies have the same chance of ending up in a linker-tag molecule, the underlying statistics of comparing SAGE experiments is also simple. The sampling of specific tags by sequencing can then be considered as sampling with replacement [41]. Therefore, the identification of differentially regulated genes between two biological samples is to test whether the observed tag counts in both samples are equal [42]. This is relatively easy and statistically well founded if a measure of variance is available. This is normally obtained by performing replicate analyses. So far, this has been the Achilles' heel in SAGE analysis since replicates of experiments have generally not been performed due to the cost of sequencing. Therefore, the necessary information on biological and technical variations has not been available. Several published statistical tests have based their tests on their own assumptions about the statistical distribution of SAGE tags from which a measure of variance is obtained [15, 43–45]. With the exception of the test proposed by Madden *et al.* [45], which is more conservative, they tend to lead to very similar results. However, with the sequencing power available today and ease of the DeepSAGE protocol, biological replicates can and should be included in the experimental design to provide a true measure of variance.

One issue, however, directly influencing the underlying assumption that every mRNA copy has the same chance of ending up in a linker-tag molecule, was introduced with the change from SAGE, using *BsmFI* [18], to LongSAGE, using *MmeI* [36]. *BsmFI* produces blunt-ended tags, where every tag can ligate together to any other tag. In contrast, LongSAGE and DeepSAGE use *MmeI* to sticky ended tags. Therefore, theoretically an observation of a tag can be suppressed by the unavailability of sufficient complementary overhangs. However, a careful examination of data sets, including the extreme expression profiles of highly secreting tissue such as pancreas, could not detect any bias in the gene expression profiles [25]. Nonetheless, such a bias cannot be entirely ruled out, when the expression profiling of smaller and smaller samples, and therefore potentially more and more specialized cell populations, becomes possible. One way of circumventing this problem is to perform the DeepSAGE analysis by using monotags (see Figure 17.1, DeepSAGE (Solexa)), where surplus of any overhang is supplied in the form of Linker-B.

Figure 17.2 Example of DeepSAGE data from potato tuber [23]. The data have been mapped to the potato EST database STGI v. 10 at <http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=potato>. Note that even abundant tags may not have precise match in the gene database, reflecting that the current

mapping database is incomplete, especially at the 3' end of transcripts, and is composed of several different cultivars of potato. SAGE studies on organisms for which a full genome sequence is available (e.g., human) have fewer unknown tags.

A very real problem when performing transcriptome profiling is the accumulation of type I statistical error: false positives. Consider a comparison of two gene expression profiles, each containing the same 10 000 different tags. This would result in 10 000 comparisons between tag pairs. Setting the significance value (p) at 0.05 for 95% confidence that the tag counts in question are not equal for each comparison would result in approximately 500 false positives. This hampers data interpretation and overwhelms experimental setups for the validation of candidate genes. Therefore, it is important to control the rate of false positives. This can be done in several ways including Bonferroni correction [46] or controlling the false discovery rate (FDR) [47]. We find it useful to make the initial comparison using a strict Bonferroni correction, which in the above example is equivalent of setting the threshold p -value at 0.000 005, reflecting that in the resulting list of differentially expressed genes, there is only a 5% chance of a single false positive. In a typical SAGE study encompassing 50 000 tags, 100–200 genes are identified as differentially regulated between two samples using this threshold. Following manual interpretation of this set, support for identified candidate pathways is found by searching the much larger list (500–1000 genes) of potentially differentially regulated genes obtained without any correction for multiple testing. Again, scripts for performing these comparisons with and without Bonferroni correction can be found at http://www.bio.aau.dk/en/biotechnology/software_applications/perl_scripts.htm.

17.5

Future Perspectives

It is known that many transcripts (e.g., transcription factors) require very low copy numbers to have a biological effect. These functionally relevant transcripts can only be reliably detected if the sensitivity of the transcriptome profiling methods is increased to include the detection of a much greater, if not the entire, part of transcriptome of a cell. This is only possible with scalable tag-based transcriptome analysis methods such as DeepSAGE, where the cost of sequencing is essentially the only limitation to sampling depth.

Therefore, it is important to use the increased sequencing power not only to process many more biological samples and include replicates in the experimental design but also to probe the transcriptome much deeper. A more comprehensive analysis of the transcriptome will provide the dynamics of the fixed genome sequence and serve as the fundament of modeling complex organisms by systems biology methods.

Acknowledgment

This work has been supported by the Danish Veterinarian and Agricultural Research Council (23-02-0034).

References

- 1 Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., ColladoVides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
- 2 Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Life with 6000 genes. *Science*, **274**, 563–567.
- 3 Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- 4 Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R.L., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchinson, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J.P., Miguel, T., Paszkowski, U., Zhang, S.P., Colbert, M., Sun, W.L., Chen, L.L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y.S., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalima, T., Oliphant, A. and Briggs, S. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science*, **296**, 92–100.
- 5 Yu, J., Hu, S.N., Wang, J., Wong, G.K.S., Li, S.G., Liu, B., Deng, Y.J., Dai, L., Zhou, Y., Zhang, X.Q., Cao, M.L., Liu, J., Sun, J.D., Tang, J.B., Chen, Y.J., Huang, X.B., Lin, W., Ye, C., Tong, W., Cong, L.J., Geng, J.N., Han, Y.J., Li, L., Li, W., Hu, G.Q., Huang, X.G., Li, W.J., Li, J., Liu, Z.W., Li, L., Liu, J.P., Qi, Q.H., Liu, J.S., Li, L., Li, T., Wang, X.G., Lu, H., Wu, T.T., Zhu, M., Ni, P.X., Han, H., Dong, W., Ren, X.Y., Feng, X.L., Cui, P., Li, X.R., Wang, H., Xu, X., Zhai, W.X., Xu, Z., Zhang, J.S., He, S.J., Zhang, J.G., Xu, J.C., Zhang, K.L., Zheng, X.W., Dong, J.H., Zeng, W.Y., Tao, L., Ye, J., Tan, J., Ren, X.D., Chen, X.W., He, J., Liu, D.F., Tian, W., Tian, C.G., Xia, H.G., Bao, Q.Y., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W.M., Li, P., Chen, W., Wang, X.D., Zhang, Y., Hu, J.F., Wang, J., Liu, S., Yang, J., Zhang, G.Y., Xiong, Y.Q., Li, Z.J., Mao, L., Zhou, C.S., Zhu, Z., Chen, R.S., Hao, B.L., Zheng, W.M., Chen, S.Y., Guo, W., Li, G.J., Liu, S.Q., Tao, M., Wang, J., Zhu, L.H., Yuan, L.P. and Yang, H.M. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). *Science*, **296**, 79–92.
- 6 Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyra, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Graves, E.D., Gregory, S., Guigo, R., Guyer, M.,

- Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., LeVine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., Mclay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Larty, E.O., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Strange-Thomann, N.S., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Vidal, A.U., Vinson, J.P., von Niederhausern, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C. and Lander, E.S. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
- 7 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W.,

- Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramsier, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la, B.M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de, J.P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S. and Chen, Y.J. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- 8 Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., bu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di, F.V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Rombold, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D.,

- Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N. and Nodell, M. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
- 9 Ruan, Y.J., Le Ber, P., Ng, H.H. and Liu, E.T. (2004) Interrogating the transcriptome. *Trends in Biotechnology*, **22**, 23–30.
- 10 Brady, S.M., Long, T.A. and Benfey, P.N. (2006) Unraveling the dynamic transcriptome. *The Plant Cell*, **18**, 2101–2111.
- 11 Siddiqui, A.S., Khattra, J., Delaney, A.D., Zhao, Y., Astell, C., Asano, J., Babakaiff, R., Barber, S., Beland, J., Bohacec, S., Brown-John, M., Chand, S., Charest, D., Charters, A.M., Cullum, R., Dhalla, N., Featherstone, R., Gerhard, D.S., Hoffman, B., Holt, R.A., Hou, J., Kuo, B.Y., Lee, L.L., Lee, S., Leung, D., Ma, K., Matsuo, C., Mayo, M., McDonald, H., Prabhu, A.L., Pandoh, P., Riggins, G.J., de Algara, T.R., Rupert, J.L., Smailus, D., Stott, J., Tsai, M., Varhol, R., Vrljicak, P., Wong, D., Wu, M.K., Xie, Y.Y., Yang, G., Zhang, I., Hirst, M., Jones, S.J., Helgason, C.D., Simpson, E.M., Hoodless, P.A. and Marra, M.A. (2005) A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 18485–18490.
- 12 Alwine, J.C., Kemp, D.J. and Stark, G.R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5350–5354.
- 13 Adams, M.D., Soares, M.B., Kerlavage, A.R., Fields, C. and Venter, J.C. (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genetics*, **4**, 373–386.
- 14 Crookshanks, M., Emmersen, J., Welinder, K.G. and Nielsen, K.L. (2001) The potato tuber transcriptome: analysis of 6077 expressed sequence tags. *FEBS Letters*, **506**, 123–126.
- 15 Audic, S. and Claverie, J.M. (1997) The significance of digital gene expression profiles. *Genome Research*, **7**, 986–995.
- 16 Ohlrogge, J. and Benning, C. (2000) Unraveling plant metabolism by EST analysis. *Current Opinion in Plant Biology*, **3**, 224–228.
- 17 Lockhart, D.J., Dong, H.L., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C.W., Kobayashi, M., Horton, H. and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675–1680.
- 18 Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
- 19 Mockler, T.C. and Ecker, J.R. (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**, 1–15.
- 20 Nielsen, K.L., Grønkjær, K., Welinder, K.G. and Emmersen, J. (2005) Global transcript profiling of potato tuber using LongSAGE. *Plant Biotechnology Journal*, **3**, 175–185.
- 21 Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R.,

- Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- 22 Bentley, D.R. (2006) Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, **16**, 545–552.
- 23 Nielsen, K.L., Hogh, A.L. and Emmersen, J. (2006) DeepSAGE: digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Research*, **34**, e133.
- 24 Heidenblut, A.M., Luttges, J., Buchholz, M., Heinitz, C., Emmersen, J., Nielsen, K.L., Schreiter, P., Souquet, M., Nowacki, S., Herbrand, U., Kloppel, G., Schmiegel, W., Gress, T. and Hahn, S.A. (2004) aRNA-longSAGE: a new approach to generate SAGE libraries from microdissected cells. *Nucleic Acids Research*, **32**, e131.
- 25 Emmersen, J., Heidenblut, A.M., Hogh, A.L., Hahn, S.A., Welinder, K.G. and Nielsen, K.L. (2007) Discarding duplicate ditags in LongSAGE analysis may introduce significant error. *BMC Bioinformatics*, **8**, 92.
- 26 Gowda, M., Jantasuriyarat, C., Dean, R.A. and Wang, G.L. (2004) Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant Physiology*, **134**, 890–897.
- 27 Lu, J., Lal, A., Merriman, B., Nelson, S. and Riggins, G. (2004) A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics*, **84**, 631–636.
- 28 Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, **18**, 630–634.
- 29 Huang, J., Hao, P., Zhang, Y.L., Deng, F.X., Deng, Q., Hong, Y., Wang, X.W., Wang, Y., Li, T.T., Zhang, X.G., Li, Y.X., Yang, P.Y., Wang, H.Y. and Han, Z.G. (2007) Discovering multiple transcripts of human hepatocytes using massively parallel signature sequencing (MPSS). *BMC Genomics*, **8**, 207.
- 30 Gowda, M., Venu, R.C., Raghupathy, M.B., Nobuta, K., Li, H.M., Wing, R., Stahlberg, E., Coughlan, S., Haudenschild, C.D., Dean, R., Nahm, B.H., Meyers, B.C. and Wang, G.L. (2006) Deep and comparative analysis of the mycelium and appressorium transcriptomes of *Magnaporthe grisea* using MPSS, RL-SAGE, and oligoarray methods. *BMC Genomics*, **7**, 310.
- 31 Nakano, M., Nobuta, K., Vemmaraju, K., Tej, S.S., Skogen, J.W. and Meyers, B.C. (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Research*, **34**, D731–D735.
- 32 Nobuta, K., Venu, R.C., Lu, C., Belo, A., Vemmaraju, K., Kulkarni, K., Wang, W.Z., Pillay, M., Green, P.J., Wang, G.L. and Meyers, B.C. (2007) An expression atlas of rice mRNAs and small RNAs. *Nature Biotechnology*, **25**, 473–477.
- 33 Meyers, B.C., Lee, D.K., Vu, T.H., Tej, S.S., Edberg, S.B., Matvienko, M. and Tindell, L.D. (2004) Arabidopsis MPSS: an online resource for quantitative expression analysis. *Plant Physiology*, **135**, 801–813.
- 34 Hene, L., Sreenu, V.B., Vuong, M.T., Abidi, S.H., Sutton, J.K., Rowland-Jones, S.L., Davis, S.J. and Evans, E.J. (2007) Deep analysis of cellular transcriptomes: LongSAGE versus classic MPSS. *BMC Genomics*, **8**, 333.

- 35 Pleasance, E.D., Marra, M.A. and Jones, S.J.M. (2003) Assessment of SAGE in transcript identification. *Genome Research*, **13**, 1203–1215.
- 36 Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002) Using the transcriptome to annotate the genome. *Nature Biotechnology*, **20**, 508–512.
- 37 Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Jr, Hieter, P., Vogelstein, B. and Kinzler, K.W. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.
- 38 Kang, J.J., Watson, R.M., Fisher, M.E., Higuchi, R., Gelfand, D.H. and Holland, M.J. (2000) Transcript quantitation in total yeast cellular RNA using kinetic PCR. *Nucleic Acids Research*, **28**, e2.
- 39 Anisimov, S.V., Tarasov, K.V., Stern, M.D., Lakatta, E.G. and Boheler, K.R. (2002) A quantitative and validated SAGE transcriptome reference for adult mouse heart. *Genomics*, **80**, 213–222.
- 40 van, R.F., Ruijter, J.M., Schaaf, G.J., Asgharnegad, L., Zwiijnenburg, D.A., Kool, M. and Baas, F. (2005) Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC Genomics*, **6**, 91.
- 41 Stollberg, J., Urschitz, J., Urban, Z. and Boyd, C.D. (2000) A quantitative evaluation of SAGE. *Genome Research*, **10**, 1241–1248.
- 42 Ruijter, J.M., van Kampen, A.H.C. and Baas, F. (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiological Genomics*, **11**, 37–44.
- 43 Kal, A.J., van Zonneveld, A.J., Benes, V., van den Berg, M., Koerkamp, M.G., Albermann, K., Strack, N., Ruijter, J.M., Richter, A., Dujon, B., Ansorge, W. and Tabak, H.F. (1999) Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Molecular Biology of the Cell*, **10**, 1859–1872.
- 44 Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B. and Kinzler, K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.
- 45 Madden, S.L., Galella, E.A., Zhu, J.S., Bertelsen, A.H. and Beaudry, G.A. (1997) SAGE transcript profiles for p53-dependent growth regulation. *Oncogene*, **15**, 1079–1085.
- 46 Altman, D. (1991) *Practical Statistics for Medical Research*, Chapman-Hall, London.
- 47 Hochberg, Y. and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Statistics in Medicine*, **9**, 811–818.

18

The New Genomics and Personal Genome Information: Ethical Issues

Jeantine E. Lunshof

18.1

The New Genomics and Personal Genome Information: Ethical Issues

Do developments in the new genomics, its applications, and technologies raise genuinely new ethical issues? Has biomedical ethics reached its boundaries? We must ask, whether novel concepts are needed to keep pace with the rapid developments in genomics, or whether we will be able to move the boundaries and provide answers within our current ethical framework.

The answers to these and other key questions depend on the interpretation of the role of ethics, in particular, as applied to the biomedical sciences. One possible interpretation, the one that I subscribe to, is expressed in the view that “Ethical thinking will inevitably continue to evolve as the science does . . .,” as voiced by Knoppers and Chadwick [1] in their landmark study.

The need for a revision of the approach taken by biomedical ethics with regard to questions concerning genomics has been appreciated for years, but developments in ethics are slow compared to the dynamic growth of genomics research [2, 3].

In this chapter, I will first identify the features of the new genomics that pose special challenges to ethics. Then I will outline the development and the structure of the current framework of mainstream biomedical ethics. From that context, the ethical issues surrounding “new style” personal genome information will be taken as an example to explain the emerging questions and to explore possible solutions.

18.2

The New Genomics: What Makes it Special?

What is Special about the New, Post-Human Genome Project Genomics’ Research? At least four striking features can be mentioned that have relevant effect on the normative and governance structures, as commonly used for dealing with genomics by society:

- *Scale and pace*: For example, the consortium efforts and networks of networks involved in genome-wide association studies (GWAS). The magnitude of the studies allows outcomes to be obtained much faster and provides the power needed for finding genome/phenome associations that would not otherwise be attainable.
- *Technology and methods*: Notably, the advances in high-throughput sequencing, but also, for example, array technology, multiplex PCR, and bioinformatics.
- *Theory development and shifting paradigms*: Hypothesis-free and hypothesis-generating research (GWAS), shift toward systems biology, and bioinformatics.
- *Novel practices*: Among other things, centralized data storage, data accessibility and sharing, and powerful web-based search engines.

A special and highly relevant feature of these developments is that we are witnessing very large-scale research, in particular the consortium efforts in genome-wide association studies, and very small-scale yet comprehensive research of individual whole genome sequencing in the form of “personal genomes” at the same time. Both applications are enabled by the new sequencing technologies and raise a specific set of ethical questions. Further advances in, for example, functional genomics, transcriptomics, epigenomics and bioinformatics, and mining of a very large number of health records will increase the information yield from the resulting comprehensive data sets, and this will also influence the potential ethical implications.

18.3

Innovation in Ethics: Why do We Need it?

The features mentioned above suggest that today’s biomedical research ethics may not be fully adequate to deal with the questions that are at stake in the post-Human Genome Project (HGP) era that presently is marked by, among other things, next-generation sequencing and gene expression technologies. There is a pressing need to address questions that span from very large-scale research projects to single personal genomes. Given the dynamics of the field, we should be prepared to face more new questions and challenges soon.

To make clear the necessity of innovative solutions, we must look at the way in which biomedical ethics has developed till now. Medical ethics has come a long way since the ancient times of Hippocrates. Although this may seem very remote from the topic of new sequencing technologies and the like, we will see that the “Hippocratic ideal,” as I would like to call it, influences the image of medical confidentiality until today, and this needs to be taken into account when, for example, redesigning consent for participation in studies in the new genomics.

The following section offers a short overview of the development of mainstream biomedical ethics, from doctor–patient-focused clinical ethics to research ethics, including epidemiological and other research work with groups and populations as

subjects. It will become apparent that the new genomics and personal genomes do raise some genuinely new questions that challenge the boundaries of the current normative frameworks.

18.4

A Proviso: Global Genomics and Local Ethics

One important proviso needs to be made. Contemporary Western world biomedical ethics is taken here as the reference. This should not be interpreted as a lack of awareness of the existence of important traditions of medical ethics in many parts of the world. Actually, the issue of how the ethical and legal norms that make up local normative frameworks can be incorporated into the governance of research and application of globalized genomics is a major research topic in the social and political sciences [4].

Mainstream Western world biomedical ethics is deliberately chosen here as a starting point because of its dominant role in framing regulation and guidelines for research worldwide, as well as the predominance of Western culture, values, and resources in funding and carrying out the majority of current human clinical genetics research. When we raise the question about boundaries, we mean the boundaries of this particular framework.

18.5

Medical Ethics and Hippocratic Confidentiality

Medical ethics has developed over millennia in many different cultural, medical, and religious contexts. Traditional codices may contain rules about how to behave among colleagues – medical etiquette – and how to act as a good doctor in the physician–patient relationship – medical ethics. In the Western tradition from Hippocrates to Percival and the first Code of Ethics of the American Medical Association, emphasis has been on professional virtues and, over the centuries, *beneficence* has been the core moral concept [5]. Over the centuries, the duties of the physician are fixed and made known to society in the condensed form of the physician’s pledge, prayer, or oath, as from Maimonides or Hippocrates, which are being continued through the modern codices [6]. Of particular relevance to our topic of personal genome information is the universal traditional promise of *strict confidentiality* in the patient–physician relationship. The Oath of Hippocrates entails the paradigmatic commitment to professional secrecy that has been generally acknowledged as a key feature of the practice of medicine ever since.

It is very important to note that this “Hippocratic ideal” of confidentiality and secrecy is commonly present in peoples’ minds today, even if the evidence suggests otherwise [7]. One illustration of the purposeful use of this omnipresent association is the fact that IBM chose the name “Hippocratic Database” for its innovative

database technology that is claimed to enable secure handling of electronic health records [8].

18.6

Principles of Biomedical Ethics

When considering modern medical ethics, we should never forget that our current twentieth-century framework of biomedical ethics arose as a reaction to a dark past of crimes against humanity, of scandals, and of medical malpractice [9, 10]. Therefore, it is no surprise that *protection* is the key concept, and the leading paradigm of biomedical ethics is the *protection paradigm*. The prime object of protection is *human dignity* [11]. Among the values derived from this core notion are the value of life and bodily integrity of individuals, the individual free will, and the acknowledgment of the right to self-determination [12, 13]. Therefore, *autonomy* is the central concept. The requirement of *respect for autonomy* has become the leading principle in postwar Western world ethics [14].

The classical principles of biomedical ethics consist of the well-known four clusters of norms that were first presented by Beauchamp and Childress in their seminal textbook, *Principles of Biomedical Ethics*, in 1979 [15]:

1. *Respect for autonomy*: Respecting the decision-making capacities of autonomous persons, this establishes the requirements of voluntariness and of consent.
2. *Nonmaleficence*: Avoiding the causation of harm.
3. *Beneficence*: Norms for providing benefits and balancing benefits against risks and costs.
4. *Justice*: Norms for distributing benefits, risks, and costs fairly.

These principles are guiding the relationship between the individual patient and the physician in clinical practice, as well as the relationships in clinical research.

18.7

Clinical Research and Informed Consent

In the context of clinical research, this model underpins the fundamental requirements of informed consent, voluntary participation free of undue constraints, a favorable risk–benefit ratio for participants, and the right to withdraw at any time from a study without consequences for medical care, to mention the most prominent ones [11]. Ethical review of whether these requirements are met and Institutional Review Board (IRB) approval of study protocols are the central procedural elements in any type of research. To protect the research subjects from harm that results from research participation is the first and foremost aim of the whole procedure. However, even the most meticulous ethics review cannot prevent

serious or even fatal harm from occurring, as recent clinical research tragedies have shown [16].

18.8

Large-Scale Research Ethics: New Concepts

Knoppers and Chadwick have presented arguments for the need of a shift in emphasis of the principles that make up the normative framework of research ethics, with focus on research in human genetics in particular [1]. Such a shift in emphasis of ethical principles does not imply disqualifying or discarding the old ones, nor does it mean giving up core moral values. As the authors say, “There might not, and cannot, be universal norms in bioethics, as emerging ethical norms are as ‘epigenetic’ as the science they circumscribe” [1]. Indeed, the issue of the universality of norms touches upon the most fundamental questions of ethics that will always be hotly contested and cannot be, once and for all, resolved by rational discourse [17].

The completion of the HGP has confronted us with a shift in emphasis in genomics research that can be characterized as a shift from individual and family disease-oriented clinical genetics – dominant in the early and mid-1990s – to population-directed research genetics, carried out in projects of formerly unknown dimensions and including populations and researchers on a global scale. What can be criteria to assess the ethical quality of studies targeting groups, communities, or populations? A novel set of ethical principles, very distinct from the well-known four principles and relevant to the rights and interests of groups, has been proposed. In keywords,

- reciprocity
- mutuality
- solidarity
- citizenry
- universality.

Obviously, the proposed framework is only a first step, and the potential impact and best way of applying each of these principles is not clear yet. Likely, it will be most suitable at the level of research planning and policy making, in particular with regard to biobanking and research directed at human genetic variation, as for example, the HapMap consortium. First results from HapMap II are available, and more distinct population subgroups will be included in ongoing research [18]. The availability and accessibility of comprehensive information from identified groups and communities raise questions that cannot be sufficiently dealt with by individual-oriented biomedical ethics. At this point, the proposed set of principles is crossing the boundaries of traditional ethics. Yet, very large-scale databases, after all, consist of individual data sets. The position of the individuals that make up the group remains unclear. Advancing technologies bring a rapid increase in scale, pace, and yield of information

content, thereby closing the gap between “anonymous” population research and personal genomics: both are “open” in the end.

18.9

Personal Genomes

No doubt, the year 2007 will be remembered, even beyond the biomedical research community, as the year of the personal genomes [19, 20], when the first personal whole genome sequences became available, one of them being the first human diploid genome [21, 22]. However, for both individuals so far only genotype data are publicly available and accessible at the NCBI Trace Archive [23]. In Summer 2007, at Harvard Medical School, the official launch of the first phase of the Personal Genome Project (PGP) took place in which 10 volunteers are involved, among them the PI of the project, George Church. At the next stage, a further expansion of the project is planned that may at some point in future include more than 100 000 volunteers [20].

18.9.1

What is a Personal Genome and What is New About It?

The term “personal genome” refers to a comprehensive genotype plus phenotype data set that, preferably, also contains information on environmental exposure and nutrition. A personal genome is by definition identifying and there is no way of ruling this out. Taking into account that any tiny piece of DNA information is identifying, it is obvious that any additional data reveal more about the individual they stem from. The American Society of Human Genetics says in its statement on GWAS:

[Being] acutely aware that the most accurate individual identifier is the DNA sequence itself or its surrogate here, genotypes across the genome. It is clear that these available genotypes alone, available on tens to hundreds of thousands of individuals in the repository, are more accurate identifiers than demographic variables alone; the combination is an accurate and unique identifier [24].

This sets the scene for delineating the challenges to current ethical practices, in particular to the practice of promising privacy and confidentiality as a condition for obtaining consent.

Samples containing DNA are available in repositories since long, well before the advent of genome-wide association studies or the concept of personal genomes. Every clinical pathology collection constitutes a biobank and in combination with data from, for example, the medical records, “personal genomes” could easily be derived. However, such collections have quite different purposes. If people give consent, assurances of strict maintenance of confidentiality are among the standard conditions of the consent forms.

The same applies to the clinical research setting where voluntariness, consent, and a favorable risk–benefit ratio to study participants are essential criteria. Promises of

protecting privacy and maintaining confidentiality – up to complete anonymity – is the rule, any deviation from it needs thorough justification. Individual research participants will likely have an image in mind that differs considerably from the realities of biomedical research, where, for example, data sharing is part of good research practice and required by oversight bodies and funding agencies [25].

Complex procedural and statistical measures are employed to keep the data deidentified. Goals are protecting the privacy and confidentiality, respecting the autonomy of the sample donors, and protecting them from harm that could arise from use of the data. The strategy is in full accordance with the requirements of the protection paradigm of modern biomedical ethics.

18.9.2

But, Can Making Promises that Cannot be Substantiated be Ever Morally Justifiable?

Turning toward the reality of dealing with data having rich information content, in an increasingly wired world, we cannot ignore that “Hippocratic confidentiality” is merely an ideal image, which even in face-to-face clinical practice no longer exists.

Efforts for privacy protection can be made, but even when they work well strict confidentiality and anonymity cannot be guaranteed. Examples of violation of privacy and breaches of confidentiality are abundant, and occur in spite of data protection measures in all areas of modern life [26]. Moreover, there is increasing evidence from both fields of medical informatics and statistics that even the most advanced anonymization techniques are vulnerable to attacks. The recently developed strategy of k-anonymization that is used in, for example, IBMs “Hippocratic Database,” has already been challenged by the technique of so-called l-diversity, which is claimed to be more robust [27, 28]. But the next attack is likely around the corner and we are certainly on the safe side when assuming that anonymization is impossible. We should therefore refrain from promises and making confidentiality a condition in consent.

Sacrificing the promise of confidentiality seems like giving up the moral foundations of biomedical research. At this point, the boundaries of the established ethical frameworks are about to be transgressed.

18.10

The Personal Genome Project: Consenting to Disclosure

The basic assumption of the Personal Genome Project that was developed alongside with the 2003 Harvard CEGS–MGIC proposal aiming at ultralow-cost/high-precision genomic imaging technology is that maximum comprehensive genotype–phenotype data sets are needed to obtain meaningful results in hypothesis-free, systems biology-based research [29, 30].

Fully consented data sets are needed from volunteer participants to make this research justified. Taking into account that secure anonymization is not possible,

the only consistent conclusion is that full and valid consent implies the abandonment of any of the conventional “confidentiality” clauses in the informed consent procedure.

In the “open consent,” as designed for the PGP, *veracity* has been determined as the lead principle. Veracity is a necessary, though not sufficient, condition for autonomy and thus for valid consent. The PGP remains, in this respect, within the boundaries of the established moral concepts.

But, what does it mean to abandon the language of confidentiality clauses? What is it that prospective research participants positively consent to? At this point, open consent implies consenting to disclosure. Volunteers agree with full and public disclosure of their comprehensive data set, consisting of genome sequence data and extensive phenotype information, including data from personal health records and facial photographs. The comprehensive data sets may or may not be made publicly accessible.

In choosing this novel consent model, the Personal Genome Project clearly moves beyond the boundaries of established practices in research ethics. The concept will further evolve, as the next-generation sequencing technologies will do. One of the great assets of the open consent protocol is that it is open to all to watch and comment upon.

Acknowledgments

I wish to thank the colleagues from the Genes Without Borders project team for ongoing inspiring discussion. I owe them many of my insights on relevant aspects of global genomic governance. I am grateful to the GEN-AU (Genomeresearch in Austria) program of the Federal Austrian Ministry of Science and Research for enabling my participation in the Genes Without Borders project.

The views expressed in this chapter are entirely my own.

References

- 1 Knoppers, B.M. and Chadwick, R. (2005) Human genetic research: emerging trends in ethics. *Nature Reviews. Genetics*, **6**, 75–79.
- 2 Editorial: Defining a new bioethic (2001) *Nature Genetics*, **28**, 297–298.
- 3 Chadwick, R. and Berg, K. (2001) Solidarity and equity: new ethical frameworks for genetic databases. *Nature Reviews. Genetics*, **2**, 318–321.
- 4 Genes Without Borders: Towards Global Genomic Governance. A project of the “Transformation in Public Policy” research group, Department of Political Analysis, University of Vienna, Austria. <http://www.univie.ac.at/transformation/GwB> (Accessed 6 November 2007).
- 5 Faden, R.R. and Beauchamp, T.L. (1986) *A History and Theory of Informed Consent*, Oxford University Press, Oxford, pp. 53–63.
- 6 World Medical Association (2006) *World Medical Association International Code of Ethics* (London, England, 1949). Adopted by the General Assembly (Pilanesberg, South Africa). <http://www.wma.net/e/>

- policy/c8.htm (Accessed 6 November 2007).
- 7 Carman, D. and Britten, N. (1995) Confidentiality of medical records: the patient's perspective. *British Journal of General Practice*, **45**, 485–488.
 - 8 Agrawal, R. and Johnson, C. (2007) Securing electronic health records without impeding the flow of information. *International Journal of Medical Informatics*, **76**, 471–479.
 - 9 Nuremberg Code: Trials of War Criminals Before the Nuremberg Military Tribunals Under Control Council Law (1949) U.S. Government Printing Office, Washington, DC, USA, pp. 181–182.
 - 10 The Belmont Report (1979) National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research, DHEW Publications, Washington, DC, USA.
 - 11 The World Medical Association (2000) The Declaration of Helsinki. World Medical Association (1964), Edinburgh. <http://www.wma.net>.
 - 12 UNESCO (2005) Universal Declaration on Bioethics and Human Rights. UNESCO, Paris. <http://portal.unesco.org/>.
 - 13 Häyry, M. and Takala, T. (2005) Human dignity, bioethics and human rights. *Developing World Bioethics*, **5**, 225–233.
 - 14 Gillon, R. (2003) Ethics needs principles – four can encompass the rest – and respect for autonomy should be “first among equals.” *Journal of Medical Ethics*, **29**, 307–312.
 - 15 Beauchamp, T.L. and Childress, J.F. (2008) *Principles of Biomedical Ethics*. 6th edn. Oxford University Press, Oxford.
 - 16 Wood, A.J.J. and Darbyshire, J. (2006) Injury to research volunteers: the clinical-research nightmare. *The New England Journal of Medicine*, **354**, 1869–1871.
 - 17 Beauchamp, T.L. (1991) *Philosophical Ethics: An Introduction to Moral Philosophy*. 2nd edn. McGraw-Hill, New York.
 - 18 The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–862.
 - 19 Check, E. (2007) Celebrity genomes alarm researchers. *Nature*, **447**, 358–359.
 - 20 Blow, N. (2007) The personal side of genomics. *Nature*, **449**, 627–630.
 - 21 SoRelle, R. Nobel laureate James Watson receives personal genome in ceremony at Baylor College of Medicine. *From the Labs*, **6** (5) (2007) <http://www.bcm.edu/fromthe/lab/vol06/is5/0607-1.html> (Accessed 6 November 2007).
 - 22 Levy, S. Sutton, G. Ng, P.C. Feuk, L. Halpern, A.L. Walenz, B.P. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biology*, **5** (10), e254, 10.1371/journal.pbio.0050254.
 - 23 GenBank National Center for Biotechnology Information Trace Archive. <http://www.ncbi.nlm.nih.gov/Traces/trace.cgi> (Accessed 6 November 2007).
 - 24 The American Society of Human Genetics (2006) ASHG Response to NIH on Genome-Wide Association Studies. http://www.ashg.org/pages/statement_nov3006.shtml (accessed 22 June 2008).
 - 25 National Institutes of Health (2003) Final NIH Statement on Sharing Research Data. Available from <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html> (Accessed 6 November 2007).
 - 26 The Personal Genome Project. Are Guarantees of Genome Anonymity Realistic? <http://arep.med.harvard.edu/PGP/anon.htm> (Accessed 6 November 2007).
 - 27 Sweeney, L. (2002) Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, **10**, 571–588.
 - 28 Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. (2007) l – Diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, **1**, 1, Article 3 (March 2007) <http://doi.acm.org/10.1145/1217299.1217302>.

- 29** The Personal Genome Project.
<http://www.personalgenomes.org>
(Accessed 15 January 2008).
- 30** Molecular and Genomic Imaging Center
(2003) Specialized Center of Excellence in
Genomic Science (CEGS) P50 Proposal.
[http://arep.med.harvard.edu/P50_03/
Church03.doc](http://arep.med.harvard.edu/P50_03/Church03.doc) (Accessed 6 November
2007).

Index

a

acceptor-tagged nucleotide 99
acrylamide-based copolymers 154
acrylamide gel 59
– matrix 61
– monomer 58
affymetrix genome-wide human SNP array 206
agarose networks 156
allele-specific color tagging 125
amplicon library method 45
amplicon variant analyzer software 48
anatomically modern human (AMH) 191
Arabidopsis thaliana transcriptome 221
array technology 205

b

BAC fingerprinting 113
BAC libraries 50
bacterial cells 58
bacterial colonies 58
basic local alignment search tool (BLAST) 210
– algorithm 210
– method 81
– score 84
BCR-ABL translocation 171
bead-or filter-based methods 195
bead purification 31
BEAMing method 61, 64
biological systems 3
biological warfare agents 98
biomedical ethics 245, 248
bisulfite-treated DNA 211, 212
blunt-ended tags 237

c

Caenorhabditis elegans 217
CAP adapter 30

capillary array electrophoresis (CAE) 5, 6, 110, 153
– development of 153
– instrument 38, 157
– system(s) 154, 157
CCD cameras 141
cDNA fragments 169
cDNA libraries 219
cDNA sequences 114
ChIA-PET experiments 177
chromatin immunoprecipitation (ChIP) 16, 29, 39, 202
– derived fragments 204
– experiments 206
– method 173, 174, 203, 205
– paired end ditagging 169
– reactions 39
– SAGE 204
– technique 202
chromatin immunoprecipitation sequencing (ChIP-seq) 26, 201, 202, 208, 209
– challenges 209
– experiments 212
– history 202
– introduction 201
– medical applications 209
– method 26, 202, 203, 208, 209, 211, 212
– protocols 26, 201, 202, 208, 209
chromatin interactions 175, 176, 177
chromosomal conformation capture (3C) method 175
chromosome sequencing 72
clone-based genome sequencing 113
copy number variations (CNVs) 23
cross-hybridization noise 170
cross-linked polyacrylamide, *see* agarose networks
custom-written software 120

- cytosine 187
- hydrolytic deamination of 187

d

- data analysis, subsequent 86
- data processing software pipeline 46
- Deep SAGE 238
 - analysis experiment 231
 - data 235
 - procedure 232
 - protocol 237
- de novo* genome sequencing 36, 112
- dideoxy-based sequencing methods 207
- digital gene expression (DGE) 21
- digital karyotyping 75
- direct linear analysis (DLA) technique 120
- disease-related mutations 120
- ditags 232
- DNA (deoxy ribose nucleic acids) 15, 19, 26, 45, 104, 211
 - A-form 107
 - amplification 6, 58, 156
 - bisulfite-treated 19
 - double-stranded 107, 135
 - methylation of 211
 - sequencing of 26
 - single-stranded 105
- DNA backbone(s) 125, 143, 146
- DNA barcoding method 118, 119, 129, 131
 - DNA mapping 117
 - molecular haplotyping 117
- DNA chip 206, 230
- DNA-coated beads 59, 64
- DNA electrophoresis 157, 160
- DNA fragment(s) 5, 10, 15, 16, 18, 45, 118, 127, 128, 155, 168
 - blunt ended 16, 18
 - libraries 17, 204
- DNA hybridization technique 205, 230
- DNA-labeled beads 65
- DNA ladders 153
 - single-stranded 153
- DNA mapping method 118, 120, 121
- DNA methylation 23, 38
- DNA microarrays 170, 173, 230, 237
 - analysis 230
 - fabrication 170
- DNA miscoding lesions 187
- DNA molecular templates 134
- DNA molecules 58, 103, 107, 129, 134, 135, 137, 184, 202
- DNA polymerase(s) 4, 18, 19, 109, 119, 140, 141
 - *Bst* 140

- *Taq* 140
- *Tth* 140
- DNA polymorphisms 111
- DNA–protein interaction 173, 203, 207
- DNA replication process 99, 105
- DNA sample 127
 - electrophoretic processing of 117
 - single- molecule templates 133
- DNA separation device 5
- DNA separation matrix 159
- DNA sequence 5, 29, 34, 153
 - alignment algorithms 156
 - analysis 92, 113
 - assemblies 92
- DNA sequencing 18, 57, 97, 103, 159
 - high-throughput 57
 - libraries 31
 - platforms 167
 - real-time 97
 - technologies 43, 89, 90, 94, 111, 167
- DNASTAR 91, 92
 - solution 91–94
- DNA strand(s) 3, 4, 99, 100, 110
 - double-stranded molecules 104, 119, 130
 - synthesis 3, 4
- DNA template 4, 31, 45, 135, 147
- dual CPU Linux PC 85
- dye-labeled oligonucleotides 29, 30
- dynamic programming matrix (DPM) 80

e

- Efficient local alignment of nucleotide data (ELAND) 210
- EGFR mutations 50
- electrophoretic separation and detection 154
- end sequence profiling (ESP) 40
- enzyme-based chain-termination method 43
- enzyme-digested DNA 15
- epigenetic modifications 173
- ESP, *see* end sequence profiling
- EST, *see* expressed sequence tags
- eukaryotic cells 173
- eukaryotic genomes 50
- expressed sequence tags (ESTs) 229
 - analysis 229
 - sequencing 230

f

- false discovery rate (FDR) 238
- FASTA format data 79
- fluorescence signals 67, 230
- fluorescent tags 110
- fluorochrome intensities 146

fluorochrome-labeled nucleotides 135, 140
 fluorochrome-labeled PCR products 144
 Fourier transformation 107
 full width-at-half-maximum (FWHM) 143
 fusion transcripts 172

g

gapping nick sites 139
 Gaussian fit 145
 Gaussian-like distribution 209
 Gaussian parameters 143
 gel-based restriction mapping 130
 gel electrophoresis 16, 108, 147
 gene expression 36
 – analysis 229
 – array-based 36
 – profiles 238
 – tag-based 36
 gene Hit method, *see* SeqID Hit method
 gene identification signature (GIS) 169, 170
 genetic information 117
 – haplotype 117
 – RNA splicing pattern 117
 genome analyzer 15, 19, 20, 23, 26
 – instrument 20
 – pipeline software 20
 – software tools 21
 genome information 245–247
 genome methylation analysis 38
 genome/phenome associations 246
 genome protein/DNA interaction analysis 175
 genome resequencing 38
 genome sequencing 9, 23, 57
 – applications 23
 – epigenomics 23
 – multiplexing 26
 – protein–nucleic acid interactions 26
 – transcriptome analysis 23
 – impact of 9
 genome-tiling arrays 230
 genome-wide association studies (GWAS) 246
 genome-wide DNA interactions 209, 212
 genomic DNA fragments 119
 genomic DNA molecules 118, 135
 genomic DNA Mounting/Overlay 139
 genomic material 155
 genomic sequence assembly 120
 gigabase-sized genome 3, 50
 glass fabrication techniques 154
 GMAT, *see* ChIP-SAGE
 GS FLX system 44, 46, 50

h

Haemophilus influenzae 6
 haploid genomes 114
 haplotype analysis 114
 haplotype barcode(s) 125, 128
 heterochromatic knobs 8
 high-performance polymeric materials 160
 high-throughput DNA Sanger sequencing equipment 230
 high-throughput sequencing 246
 hippocratic database 247
 histone 24
 – ChIP-seq analysis 24
 histone-DNA interactions 203
 Hit method 82, 83
 human adenovirus 123, 124
 human/animal genome sequencing 50, 89, 114
 human diploid genome 103
 human genome project (HGP) 6, 9, 57, 117
 human transcript lengths 114
 hybridization-based gene arrays 205
 hybridization-based tag sequencing method 205
 hybridization chips 134
 hybridization probes 62
 hybridization techniques 9, 206
 hydrophobically modified polyacrylamides 159

i

Illumina's bead array technology 38
 Illumina genome analyzer II system 15–21
 – cluster creation 17–19
 – data analysis 20–21
 – library preparation 15–17
 – paired end reads 19–20
 – sequencing 19
 Illumina/Solexa and applied biosystems platforms 222
in situ hybridization 119
in situ template amplification 133
 Institute for Genomic Research 3
 Institutional Review Board (IRB) 248
 integrated microfluidic devices 154

k

Klenow fragment 188

l

lab-on-a-chip systems 156
 lambda DNA 121
 – sequence motif maps 121

- templates 148
- large-scale research ethics 249–250
- laser-induced fluorescence (LIF) 154, 156
- library preparation 15, 221
- DNA fragmentation 15–16
- end repair and ligation 16–17
- ligation-based sequencing 29
- ligation-mediated amplification (LMA) 175
- linear amplification 105
- linear DNA mapping techniques 120
- linear polyacrylamide (LPA) solution 157
- linker-flanked tags, *see* ditags
- linker-tag molecule 232, 237
- liquid-handling platforms 5
- locus control regions (LCRs) 175
- long-range PCR fragments 118
- LPA matrix 157

m

- mammalian genes 114
- mammalian transcriptomes 168
- massive parallel signature sequencing (MPSS) 66, 133, 221, 222, 234
- Mathies lab device 154
- Maxam–Gilbert's method 10
- microarray technologies 39
- microchip systems 154
- microfluidic devices 135, 153, 154, 156, 160, 231
- Agilent's Bioanalyzer 231
- PDMS 139
- microRNAs (miRNAs) 39, 217, 225
- background 217
- cloning frequencies of 225
- coding genes 217
- discovery 217, 223
- expression profiling 217, 225
- identification 218
- pathway model 218
- sized fragments 221
- molecular haplotyping, technology 125
- mRNA population, *see* mammalian transcriptome
- mRNA tag library sequencing 57
- mtDNA genomes 183
- *Anomalopteryx didiformis* 183
- *Dinornis giganteus* 183
- *Emeus crassus* 183
- multiplex sequencing of paired end ditags (MS-PET) 168

n

- nanosequencing machines 99
- neuron-restrictive silencer factor (NRSF) 208

- next-generation sequencing (NGS) 79, 83, 86
- data analysis 79, 86
- data sets 83
- projects 86
- sequence analysis, strategies 84
- system 27
- technologies 177
- next-generation software, DNASTAR's 89
- nicked/gapped templates 135
- Northern blot 223, 225, 230
- NRSF binding motif 208
- NRSF binding site 208
- nucleic acid sequence 133
- nucleotide pairs 104

o

- Olympus IX-71 microscope 120
- optical sequencing 133–140, 148
- microscope 137
- reactions 140
- surfaces 137
- organelle genome 192

p

- padlock probe ligation 119
- paired end ditag (PET) sequencing 167–170
- approach 170
- mapping regions 174
- methodology 168
- strategy 167
- technology 168, 179
- paleogenomic DNA 194
- paleogenomics 183
- parallel sequencing technologies 117
- personal genome project (PGP) 250
- phiX174 bacteriophage genome 3
- photopolymerized monolith 155
- plus and minus method 3
- point spread function (PSF) 120, 141
- Poisson distribution 31
- polony karyotyping 75
- polony sequencing method 57, 59, 61, 62, 66, 69, 70
- applications 57
- polony SAGE 73
- transcript characterization 73
- transcript profiling 73
- history 57–58
- system 67
- technology 57
- poly electrolyte multilayer (PEM)-modified glass surface 119
- polyacrylamide gel (PAG) 5
- polyadenylation sites (PAS) 168

- polymerase chain reaction (PCR) 10, 30, 31, 45, 58, 60, 109
 - amplicon 44, 141
 - amplification process 31, 43, 63, 108
 - amplified templates 33
 - based ultra-deep sequencing 47
 - chamber 156
 - inhibiting molecules 156
 - primers 34, 44, 124, 125, 176
 - primer sites 64
 - products 8, 143
 - reagents 58
 - sequencing analysis 184
 - polymer networks 156
 - polymorphic alleles tagged 125
 - direct haplotype determination 127
 - localization of 125
 - primer extension reactions 147
 - protein–DNA interaction 39, 201, 203, 204, 205, 209, 212
 - mapping of 201
 - sites 212
 - protein–nucleic acid interactions 21, 26
 - protein–protein interactions 203
 - prototype technique, *see* plus and minus method
- q**
- Q-PCR, *see* Northern blot
- r**
- rapid amplification of cDNA ends (RACE) reactions 230
 - RDBMS model 86
 - reaction chamber setup 137–139
 - real-time DNA sequencing 97, 100
 - real-time PCR method 62
 - repressor element-1 silencing transcription factor (REST) 208
 - restriction enzyme 235
 - *BsmFI* 235
 - *DpnII* 235
 - *NlaIII* 235
 - *Sau3A* 235
 - reverse transcription 124
 - rhinovirus genomes 123
 - ribosomal RNA band 231
 - RNA-degrading RNases 231
 - RNA-induced silencing complex (RISC) 217
 - RNA splicing patterns 130
 - RNA virus 124
- s**
- SABE, *see* serial analysis of binding elements
 - SAGE, *see* serial analysis of gene expression
 - Sanger DNA sequencing 5, 6, 10, 50, 91, 133, 153, 156
 - approaches 133
 - basics of 3
 - capillary electrophoresis 133
 - dideoxy-based tag sequencing 203, 204, 205
 - instruments 36
 - limitation and opportunities 7
 - method 4, 10, 90, 117, 153
 - principle of 4
 - metagenomic libraries 183
 - SeqID Hit method 82
 - SeqMan genome assembler (SMGA) 91
 - sequence assembly tools 21
 - sequence-based expression profiling 226
 - sequence-based techniques 40
 - sequence searching, strategies 80
 - sequence-specific probes 119
 - sequencing-by-synthesis (SBS) 133, 206
 - sequencing factories 5
 - sequencing library 63
 - sequencing technology 3, 15, 69, 109
 - ultrahigh-throughput 69
 - serial analysis of binding elements (SABE) 204
 - serial analysis of gene expression (SAGE) 113, 230
 - library 235
 - protocol 231
 - technique 204, 237
 - short-read sequencing technology 211
 - shotgun cloning 10
 - signal-to-noise (S/N) ratio 146
 - silica sol-gel monolith, *see* photopolymerized monolith
 - single fluorescent dye molecules 125
 - single-molecule sequencing technologies 110
 - single-nucleotide polymorphisms (SNPs) 3, 8, 72, 89
 - single-pair fluorescent resonance energy transfer (spFRET) 133
 - SMART technology (Clontech) 221
 - SMGA assembly projects 91
 - Smith–Waterman alignment algorithm 210
 - sodium dodecyl sulfate (SDS) 159
 - Solexa sequencing 233–235
 - SOLiD system 29, 30, 35–38, 40
 - applications 35
 - library generation 30–31
 - overview of 29
 - performance of 29, 33
 - technology of 29, 39
 - SSAHA-based methods 84

t

- tag-based sequencing 21, 36, 167
 - platforms 167
- tag-based transcriptome analysis methods 229, 238
- tag-based transcriptome profiles 235
- Tamra dye signals 121
- total internal reflectance fluorescence (TIRF) 99
 - microscopy 118
 - system 120
- transcriptional elements 202
- transcription factor-binding elements 39
- transcription factor binding sites (TFBS) 167, 168, 169, 173, 177
 - genome mapping 167, 168
- transcription regulatory circuits 173
- transcription regulatory networks 179
- transcription start sites (TSS) 168
- transcriptome 23, 229
 - high-resolution map 24
- transcriptome analysis methods 23, 37, 113, 170, 229, 230
 - DNA microarrays 230
 - serial analysis of gene expression (SAGE) 230
- transcriptome profiling 238
- transcriptome sequencing 19
- transient entanglement coupling (TEC) 158
- transmission electron microscopes (TEMs) 103
 - analysis 104
 - based sequencing 109

- image 104
- instrument 107
- sequencing technology 110
- substrate 106, 107
- technology 111
- visualization 107
- T7 exonuclease 148

u

- ultradeep sequencing 47
- ultrahigh-speed DNA sequencing 97

v

- viral genomes 123
- VisiGen's core technology 101
 - approach 100

w

- Western blotting methods 39
- whole genome shotgun (WGS) 5, 6

x

- xenon lamp 34

y

- yeast telomeric heterochromatin 204

z

- Z-labeled nucleotides (Z-dNTPs) 107
- Z-modified nucleotides (Z-dNTPs) 104, 105, 108
- ZS genetics (ZSG) 103, 111
- Z-substituted DNA molecules 103
- Z-tagged nucleotides 108